HEALTHINFO 2025

The Tenth International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing

ISBN: 978-1-68558-312-5

October 26th - 30th, 2025

Barcelona, Spain

HEALTHINFO 2025 Editors

Les Sztandera, Thomas Jefferson University, USA

Jamie McGlothlin, RSM US LLP, USA

HEALTHINFO 2025

Forward

The Tenth International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing (HEALTHINFO 2025), held between October 26th, 2025, and October 30th, 2025, in Barcelona, Spain, continued a series of events on particular aspects belonging to health informatics systems, health information, health informatics data, health informatics technologies, clinical practice and training, and wellbeing informatics in terms of existing and needed solutions.

Advances in society and technology, particularly in systems approaches, data processing, modeling, information technology, computing, and communications, have significantly improved solutions to challenges in assistive healthcare, public health, and everyday wellbeing. While achievements are tangible, open issues related to global acceptance, costs models, personalized services, record privacy, and real-time medical actions for citizens' wellbeing are still under scrutiny.

We take the opportunity to warmly thank all the members of the HEALTHINFO 2025 technical program committee, as well as all the reviewers. The creation of such a high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to HEALTHINFO 2025. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the HEALTHINFO 2025 organizing committee for their help in handling the logistics of this event.

We hope that HEALTHINFO 2025 was a successful international forum for the exchange of ideas and results between academia and industry for the promotion of progress in the field of assistive technologies for healthcare, medical support, and wellbeing.

HEALTHINFO 2025 Chairs

HEALTHINFO 2025 Steering Committee
Shada Alsalamah, King Saud University, Saudi Arabia
Nelson P. Rocha, University of Aveiro, Portugal

HEALTHINFO 2025 Publicity Chairs
Lorena Parra Boronat, Universidad Politécnica de Madrid, Spain
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
Jose Miguel Jimenez, Universitat Politecnica de Valencia, Spain

HEALTHINFO 2025
Committee

HEALTHINFO 2025 Steering Committee

Shada Alsalamah, King Saud University, Saudi Arabia
Nelson P. Rocha, University of Aveiro, Portugal

HEALTHINFO 2025 Publicity Chairs

Lorena Parra Boronat, Universidad Politécnica de Madrid, Spain
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
Jose Miguel Jimenez, Universitat Politecnica de Valencia, Spain

HEALTHINFO 2025 Technical Program Committee

Djafar Ould Abdeslam, University of Haute Alsace, France
Sherif Abdelwahed, Virginia Commonwealth University, USA
Somayeh Abedian, Ministry of Health and Medical Education, Tehran, Iran
Miriam Allalouf, Azrieli College of Engineering Jerusalem - JCE, Israel
Jens Allmer, Hochschule Ruhr West, University of Applied Sciences, Germany
João R. Almeida, University of Aveiro, Portugal / University of A Coruña, Spain
Shada Alsalamah, King Saud University, Saudi Arabia
Iqra Ameer, The Pennsylvania State University - Abington Campus, USA
Mustapha Aouache, Telecom Division - Centre de Développement des Technologies Avancées (CDTA), Algiers, Algeria
Khalfalla Awedat, Pacific LutheranUniversity, USA
Mana Azarm, University of Ottawa, Canada
Nabil Georges Badr, Higher Institute of Public Health – USJ, Beirut, Lebanon
Panagiotis D. Bamidis, Aristotle University of Thessaloniki, Greece
Hugo Barbosa, University Lusófona, Portugal
Fabio Baselice, University of Naples Parthenope, Italy
Arriel Benis, Holon Institute of Technology, Israel
Ahmed Bentajer, National School of Applied Sciences | Abdelmalek Essaad University, Tetouan, Morocco
Vilmos Bilicki, University of Szeged, Hungary
Amine Boufaied, ISITCom | University of Sousse, Tunisia
Guillaume Bouleux, University of Saint Etienne | INSA-Lyon, France
Klaus Brinker, Hamm-Lippstadt University of Applied Sciences, Germany
Tolga Çakmak, Hacettepe University, Turkey
Manuel Campos Martínez, University of Murcia, Spain
Armand Castillejo, STMicroelectronics, France
Rui Pedro Charters Lopes Rijo, Polytechnic of Leiria | INESCC | CINTESIS, Portugal
K.A.D. Chathurangika P. Kahandawaarachchi,Sri Lanka Institute of Information Technology, Sri Lanka
Sudarshan S. Chawathe, University of Maine, USA
Bhargava Chinni, University of Rochester, USA
Mario Ciampi, National Research Council of Italy - Institute for High Performance Computing and

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Discrimination by Deep Learning of 1Hz Difference in Auditory Cortex Using fMRI Activation Patterns

Yoshitaka Ooyashiki
Kochi University of Technology
Tosayamada, Kami, Kochi, 782-8502, Japan
e-mail: ooyashikiyoshitaka@gmail.com

Kyoko Shibata
Kochi University of Technology
Tosayamada, Kami, Kochi, 782-8502, Japan
e-mail: shibata.kyoko@kochi-tech.ac.jp

Abstract- **Brain decoding is a technology that interprets physical and psychological states from brain activity, and it is expected to serve as a means of medical support and communication for people with disabilities. Recently, brain decoding has gained considerable attention, especially with the advent of deep learning techniques. This study builds on the concept of tonotopy in the auditory cortex and aims to develop a method to discriminate between two sounds with a 1 Hz difference, which is difficult for humans to distinguish, using brain activation images. In a previous work, the focus was on brain activation imaging acquisition methods, and research was conducted using the two main imaging designs in functional Magnetic Resonance Imaging (fMRI) experiments: event-related design and block design. The findings indicated that both designs were effective, and further improvements in accuracy are anticipated. Therefore, this report aims to further improve discrimination accuracy. To improve accuracy, this report focused on Region of Interest (ROI) expansion, hypothesizing that an increase in activation information contributes to improved accuracy of deep learning models. In this report involved the execution of experiments in which Brodmann Areas (BA) 22 was introduced as an additional ROIs, in conjunction with the existing ROIs, BA41 and BA42. The results demonstrated that expanding the ROI improved accuracy across both designs. Notably, the block design yielded an over 30% improvement, reaching 100% discrimination accuracy. The results demonstrated that ROI expansion is an effective method for enhancing accuracy.**

Keywords- fMRI; CNN; Brain decoding; Tonotopy; Region of interest.

I. INTRODUCTION

Brain decoding is a technology that decodes physical and psychological states from brain activity, and it is a subject of extensive research in the field of neuroscience. Recently, brain decoding has gained considerable attention, especially with the advent of deep learning techniques. A substantial body of research has been dedicated to investigating the visual cortex using fMRI. For instance, there are studies such as decoding emotional expressions from visual cortex images using Support Vector Machines (SVM) and Convolutional Neural Networks (CNN) [1], and decoding objects seen in dreams from visual cortex images using Deep Neural Networks (DNN) [2]. Conversely, research in fMRI-based auditory cortex decoding has lagged that of the visual cortex. This is primarily because approximately 80% of human perceptual information is visual, prioritizing research on the visual cortex. Moreover, operational noise during fMRI scanning interferes with detecting neural activity in the auditory cortex. However, advances in fMRI technology, including the implementation of noise-canceling mechanisms, have facilitated research on the auditory cortex.

Recent auditory decoding studies have reported decoding everyday sounds from brain activity [3][4], as these are directly related to people's daily lives. Studies on decoding brain activity for music have also progressed, with reports identifying genres and moods (e.g., cheerful, somber, uplifting) from neural responses [5][6]. However, most of these studies have focused on qualitative musical characteristics such as mood, whereas studies targeting quantitative musical characteristics (e.g., frequency or sound pressure) remain scarce.

Considering early disease detection and the identification of cognitive decline, quantitative analysis is required rather than qualitative examination. Therefore, our study group is focusing on the decoding of quantitative musical characteristics. In a previous work [7], an accuracy of 75% was achieved in discriminating two sounds with a 124.5 Hz difference using deep learning applied to brain activation images. However, investigating even finer frequency differences is necessary to advance the goals of early disease detection and the identification of cognitive decline. Therefore, this study addresses the recognition of finer frequency differences to enhance frequency resolution. To investigate brain responses to auditory stimuli, the auditory cortex was defined as the ROI, following the concept of tonotopy. Tonotopy refers to the spatial organization of frequency-specific responses within the auditory cortex. This phenomenon has been confirmed in many previous studies [8][9][10][11], particularly in BA 41 and 42 (primary auditory cortex). These studies typically employed a frequency range of 1–40 kHz and focused primarily on continuous frequency modulation. However, the investigation of the potential efficacy of tonotopy functions in instances of minor frequency discrepancies remains an uncharted territory. In this study, it was motivated by the hypothesis that "even frequency differences that cannot be perceived by humans could still lead to distinct patterns of activation in the auditory cortex". In general, healthy individuals can discriminate frequency differences of about 5 Hz, while 2–3 Hz differences are influenced by musical training and individual variability. Accordingly, this study defines "imperceptible frequency differences" as a 1 Hz gap between two pure sounds. The primary aim of this study is to develop a method capable of

 ISBN: 978-1-68558-312-5

discriminating between two pure sounds differing by 1 Hz using fMRI data.

In this study, deep learning is employed to address frequency differences that are imperceptible to the human auditory system. Specifically, brain activation images are acquired while presenting two pure sounds differing by 1 Hz, and a deep learning model is used to discriminate which sound was presented based on the brain activity images. The primary challenges of this method pertain to the selection of a deep learning model and the method of acquiring brain activation images. To address the first challenge, a 3D Convolutional Neural Network (3DCNN) based model was adopted, considering the inherently three-dimensional structure of the brain. To address the second challenge, the focus will be on the two primary imaging designs employed in fMRI experiments: event-related design and block design. The event-related design is characterized by its limited capacity for image clarity due to the constrained temporal parameters allocated for imaging procedures. However, this design facilitates the acquisition of a substantial volume of data. Conversely, block design demonstrates superiority in terms of image clarity, a consequence of its prolonged imaging duration. However, this approach permits a more limited acquisition of data.

In this study, we have verified the imaging designs (event-related design and block design) that are effective for discrimination experiments using both designs. In previous works [12][13], an attempt was made to discriminate two sounds with a 1 Hz difference using deep learning, with the ROI set to BA41 and BA42 based on previous studies [8][9][10][11]. The findings indicated a discrimination accuracy of 55.90% for the event-related design and 63.41% for the block design, suggesting that both experimental designs are effective and hold promise for further enhancement of accuracy.

However, since there are no comparable previous studies for this study, we cannot rule out the possibility that accuracy is low when viewed as an absolute value of 63%. Therefore, this report aims to further improve accuracy compared to previous works [12][13]. Data augmentation has been recognized as an effective approach to enhancing accuracy. While a variety of data augmentation methods exist, the report focuses on ROI augmentation. The rationale for this focus is that the increased activation information obtained through ROI augmentation may enhance the performance of deep learning models. The proposed region for augmentation is BA22 (higher-order auditory cortex). As previously mentioned, tonotopy has been primarily confirmed in BA41 and 42; however, its presence has also been suggested in BA22 [14]. However, given the paucity of studies on tonotopy in BA22, the efficacy of BA22 in studies targeting frequency differences, as evidenced in this report, remains uncertain. Consequently, in this report, BA22 is additionally designated as a region of interest, assuming that tonotopy is also active in this area alongside BA41 and 42. In addition, given the utilization of two designs in previous works [12][13], this report employs two designs as well. This report is an individual analysis.

The structure of this report is as follows. Section II delineates the methodologies employed in brain activation imaging and frequency discrimination techniques. Section III presents the discrimination results obtained using the constructed deep learning model. Section IV investigates the factors contributing to improved discrimination accuracy and describes the discrimination techniques found to be effective based on the study's findings. Section V provides a summary.

II. METHOD

The procedure is outlined as follows. Brain activation images are obtained using an fMRI scanner while presenting two auditory stimuli differing by 1 Hz. These images are annotated using Statistical Parametric Mapping (SPM) 12 for input into deep learning. The annotated 3D data is then employed to train the model and perform discrimination using training data with a 3DCNN. A detailed explanation is provided in the subsequent section.

A. fMRI experiment

The fMRI experiment was conducted to obtain brain activation images for use in discrimination experiments. The fMRI apparatus utilized is the MAGNETOM Prisma 3T, manufactured by SIEMENS. The auditory stimuli consisted of pure sounds at 523 Hz and 524 Hz, with sound pressure levels ranging from 78 to 83 dB. Auditory stimuli were generated using Steinberg Nuendo 10.3 and delivered to participants via Opto ACTIVE thin headphones employing Active Noise Control to attenuate fMRI scanner noise. In this experiment, one 20-years-old healthy male subjects participated, who do not have abnormality in the simple hearing test. The imaging design will be a block design (Task 9 s, Rest 15 s) and an event-related design (Task 3 s, Rest 3 to 21 s in multiples of 3). This report uses brain activation images obtained in a previous work [12].

B. Annotation

The subsequent section will address the implementation of data analysis for deep learning. The conversion of the DICOM format to the NIfTI-1 format is necessary for the subsequent analysis using brain image analysis software SPM12. Subsequently, the preprocessing and individual analysis should be performed. Preprocessing included several steps: realignment to correct head motion, slice timing correction to adjust temporal differences across slices, coregistration with structural images, spatial normalization, and spatial smoothing. The objective of individual analysis is to extract brain activation characteristics through the random selection of multiple images, the implementation of statistical analysis, and the creation of contrast. In this report, regarding the training data, we created a single statistical image from two scans of brain activation images that had undergone preprocessing and obtained the following training data (per frequency) for each design by changing the combination of the two scans. There were 192 training data points for the event-related design and 80 training data points for the block design. The test data were obtained as a single statistical image from one scan, resulting in 24 test data points (per

TABLE I. NUMBER OF TRAINING DATA AND TEST DATA. LINE 1 IS EVENT-RELATED DESIGN. LINE 2 IS BLOCK DESIGN.

Designs	Training data	Test data
Event-related	192	24
Block	80	24

TABLE II. HYPER PARAMETER IN LEARNING.

Kernel size (Ks)	3, 4, 5, 6, 7
Filters (F)	16, 32
Batch size (Bs)	8, 16, 32

frequency) for each design. These are shown in Table 1. The ROIs are defined as BA41, 42, and 22, and for each contrast, the corresponding t-values and spatial coordinates are extracted. For each contrast, normalized values ranging from 0.0 to 1.0 are exported in CSV format. Using the spatial coordinates, the data are transformed into 3D arrays with dimensions H41×W50×D15 for input into the deep learning model. Voxels outside the ROI are assigned to a value of 0.0.

C. *Frequency discrimination method*

In this report, we utilize 3DCNN, a variant of deep learning, to discriminate auditory stimuli. 3DCNN represents a model that extends the capabilities of CNN, which are designed for image recognition, to three dimensions. 3DCNN utilizes convolution and pooling operations in three dimensions to extract features, thereby expanding the scope of image recognition in the third dimension. The architecture of the 3DCNN consists of sequential convolution and pooling layers, followed by a fully connected layer positioned directly before the output layer. To perform binary discrimination, the model employs two output neurons, with the softmax activation function applied to convert the outputs into probabilistic scores. The hyperparameters used for training are listed in Table 2. A grid search was performed to evaluate all possible combinations of parameter values specified in the table. Training was considered complete when the error rate dropped below the threshold of 0.1. The trained model was subsequently applied to discriminate the two auditory stimuli using test data.

III. RESULTS

The discrimination accuracy is defined as the number of correct answers obtained by inputting the test data into the trained model that has been successfully completed, divided by the total number of test data points, which is 24. A grid search was performed, and the discrimination accuracy and hyperparameters that achieved the highest accuracy after ROI expansion are shown in Table 3. Furthermore, as illustrated in Table 4, the discrimination accuracy and hyperparameters that were found to be most effective prior to ROI expansion are documented, as outlined in [13]. Table 4 presents the results for 192 training data and 24 test data.

IV. DISCUSSION

Given that this report constitutes a two-classification discrimination, the probability of a correct guess by chance is 50%, and thus the chance level is also 50%. Previous studies indicate that an accuracy exceeding 50% can be interpreted as successful discrimination [15], while an accuracy above 60%

TABLE III. HYPERPARAMETERS AND DISCRIMINATION ACCURACY IN TWO DESIGNS. (ROIS: BA41, 42, 22)

Designs	Discrimination accuracy	Hyper parameter
Event-related	60.42%	Ks:6, F:16, Bs:16
Block	100%	Ks:4,7, F:32, Bs:8

TABLE IV. HYPERPARAMETERS AND DISCRIMINATION ACCURACY IN TWO DESIGNS. (ROIS: BA41, 42) [13]

Designs	Discrimination accuracy	Hyper parameter
Event-related	55.90%	Ks:6, F:6, Bs:16
Block	63.41%	Ks:3, F:14, Bs:16

is considered sufficiently reliable [16]. As demonstrated in Tables 3 and 4, an enhancement in discrimination accuracy was observed with ROI expansion in both designs, particularly in the block design. This enhancement is likely attributable to the incorporation of activation information from BA22, which contributed to the enhancement in accuracy. Although methodological differences from a previous study [14] preclude definitive conclusions, the results suggest a potential presence of tonotopic organization in BA22.

A substantial discrepancy in the degree of discrimination accuracy enhancement was observed between the two designs. While the event-related design demonstrated a 5% enhancement in accuracy, the block design exhibited a substantial 35% improvement. This discrepancy in accuracy enhancement is hypothesized to be attributable to the clarity of the images. The disparity in image clarity between the two designs can be attributed to the following. The fMRI detects changes in cerebral blood flow induced by variations in oxygen demand following external stimuli, a phenomenon known as the BOLD effect, to acquire brain activation images. In essence, longer stimulus duration (presented for 9 s in the block design, three times longer than in the event-related design) enhances the BOLD effect, enabling clearer brain activation imaging in the block design. Thus, the notable 35% improvement in discrimination accuracy observed in the block design is considered to result from ROI expansion in clearer images, which substantially enhances the inclusion of activation-related information. In a previous work [13], the ROI was set to BA41 and BA42, the same as in [12], and accuracy was examined by doubling the training data in the block design. While this resulted in a 15% increase from approximately 48% to 63%, in absolute terms, it only achieved a marginally more reliable level of accuracy. In this report, the training data remains the same 80 data points as in [12]. However, the accuracy improvement achieved through ROI expansion is twice as much as previous work. This result demonstrates that ROI expansion yields higher accuracy than simply increasing the training data. In contrast, the lower image clarity in event-related designs likely limit the effectiveness of ROI expansion in adding activation-related information, compared to the block design. To achieve further improvements in discrimination accuracy, potential strategies include increasing the amount of training data.

Based on the results of previous works [12][13] and this report, it was determined that the following approach is effective for discriminating two tones differing by 1 Hz in brain activation images of individuals using deep learning: implementing the imaging design as a block design, creating

statistical images from two scans (as in a previous work [12]), and selecting ROIs BA41, 42, and 22. Therefore, the objective of this study was achieved. The results of this study demonstrate that discrimination within the 500 Hz band is possible and highly accurate for individuals. However, since sounds encountered in daily life exist beyond the 500 Hz band, investigation of other frequency bands is also important. Furthermore, as this study involved only one examinee and is a preliminary investigation, increasing the number of participants and verifying generalization to untrained individuals is necessary. Furthermore, the test data used in this report is limited to 24 samples. From the perspective of generalization performance in the recognition model, there is room for discussion, such as increasing the test data to verify performance.

V. CONCLUSIONS AND FUTURE WORK

Given the limited number of studies decoding quantitative musical characteristics, this study addressed the discrimination of two sounds differing by only 1 Hz a difference imperceptible to humans. The aim of this study was to improve discrimination accuracy beyond that reported in previous works [12][13]. ROI expansion was proposed as a method, and discrimination experiments were conducted using two fMRI experimental designs, event-related and block, as in a previous works [12][13]. The results showed an accuracy improvement of approximately 5% with the event-related design and over 30% with the block design. These findings demonstrate that ROI expansion is an effective approach for improving accuracy. Moreover, based on both the present results and those of previous works [12][13], it was determined that, for discriminating two sounds with a 1 Hz difference using brain activation images and deep learning, an effective strategy is to employ a block design for imaging, generate statistical images from two scans, and select ROIs in BA41, BA42, and BA22. This study provides new evidence that the brain responds even when humans are unable to perceive the difference.

Future work will include verifying generalization performance for untrained participants. With further progress, this line of this study is expected to contribute to the early detection of disease and to improvements in hearing aid performance.

REFERENCES

[1] Y. Liang, K. Bo, S. Meyyappan, and M. Ding, “Decoding fMRI data with support vector machines and deep neural networks”, Journal of Neuroscience Methods, vol. 401, pp. 110004, 2024.

[2] T. Horikawa and Y. Kamitani, “Hierarchical Neural Representation of Dreamed Objects Revealed by Brain Decoding with Deep Neural Network Features”, Frontiers in Computational Neuroscience, vol. 11, pp. 00004, 2017.

[3] M. Zhao and B. Lin, “An fMRI-based auditory decoding framework combined with convolutional neural network for predicting the semantics of real-life sounds from brain activity”, Applied Intelligence, vol. 55, pp. 118, 2025.

[4] I. R. Khan, S. L. Peng, R. Mahajan, and R. Dey, “Short-window EEG-based auditory attention decoding for neuroadaptive hearing support for smart healthcare”, Neuroscience Informatics, vol. 5, pp. 100222, 2025.

[5] M. Guilhem, M. D. L. Giovanni, and A. S. Shihab, “The Music of Silence: Part I: Responses to Musical Imagery Encode Melodic Expectations and Acoustics”, Journal of Neuroscience, vol. 41 (35), pp. 7435-7448, 2021.

[6] I. Daly, “Neural decoding of music from the EEG”, Nature, Scientific Reports. 13, pp. 624, 2023.

[7] N. Shigemoto, H. Satoh, K. Shibata, and Y. Inoue, “Study of Deep Learning for Sound Scale Decoding Technology from Human Brain Auditory Cortex”, IEEE, Global Conference on Life Sciences and Technologies, pp. 212-213, 2019.

[8] D. R. M. Langer, W. H. Backes, and P. V. Dijk, ” Representation of lateralization and tonotopy in primary versus secondary human auditory cortex”, NeuroImage, vol. 34, pp. 264-273, 2007.

[9] W. Guo et al., “Robustness of Cortical Topography across Fields, Laminae, Anesthetic States, and Neurophysiological Signal Types”, The Journal of Neuroscience, vol. 32(27), pp. 9159–9172, 2012.

[10] V. A. Kalatsky, D. B. Polley, M. M. Merzenich, C. E. Schreiner, and M. P. Stryker, “Fine functional organization of auditory cortex revealed by Fourier optical imaging”, PNAS vol. 102 No. 37, pp. 13325-13330, 2005.

[11] S. Romero et al., “Cellular and Widefield Imaging of Sound Frequency Organization in Primary and Higher Order Fields of the Mouse Auditory Cortex”, Cerebral Cortex, vol. 30, pp. 1603-1622, 2020.

[12] Y. Ooyashiki, K. Shibata, and H. Satoh, “Identification of Small Frequency Differences in Primary Auditory Cortex of Human Brain by Deep Learning Using fMRI”, JSME Chugoku-Shikoku Student Association 54th Student Member Graduation Research Presentation Lecture, pp. 01c2, 2023, (in Japanese).

[13] Y. Ooyashiki, K. Shibata, and H. Satoh, “Discrimination fMRI Learning of 1Hz Difference in Primary Auditory Cortex of Human Brain using fMRI”, JSME Mechanical Engineering Congress, pp. J162p-21, 2024, (in Japanese).

[14] J. B. Issa et al., “Multiscale optical Ca2+ imaging of tonal organization in mouse auditory cortex”, Neuron, vol. 83, pp. 944-959, 2014

[15] Thomas A. Carlson, Tijl Grootswagers, and Amanda K. Robinson;“An introduction to time-resolved decoding analysis for M/EEG”, arXiv, q-bio.NC, 1905.04820, (2019).

[16] A. K. Robinson, G. L. Quek, and T. A. Carlson, “Visual Representations: Insights from Neural Decoding”, Annual Review of Vision Science, vol. 9, pp. 313-335, 2023.

Mobility: Encouraging Physical Activity among High School Students

Maria Luiza Calisto dos Santos[1], Pedro Almir Martins de Oliveira[1 2],
Rossana Maria de Castro Andrade[2]
[1] Laboratory of Innovation and Scientific Computing (LICC)
Federal Institute of Maranhão (IFMA)
Pedreiras, Brazil
email: {luizacalisto, pedro.oliveira}@acad.ifma.edu.br
[2] Group of Computer Networks, Software Engineering and Systems (GREat)
Federal University of Ceará (UFC)
Fortaleza, Brazil
email: rossana@ufc.br

***Abstract*—With technological advancement, there is a high level of daily interaction with screens and mobile devices among the youth population, leading to a sedentary lifestyle and a reduction in the time dedicated to physical exercise. Given this context, it is essential to find solutions to address the increasing sedentary lifestyle among adolescents. In this way, this study presents the Mobility app to motivate high school students to engage in physical activities through playful physical challenges, utilizing smartphone sensors and a scoring system based on the activities completed. The goal is for students to adopt regular yet straightforward healthy habits, contributing to their overall well-being and preventing diseases associated with a sedentary lifestyle. To evaluate this proposal, a real-world evaluation was conducted with 23 students. The results show that the tool constitutes an approach that is easily integrated into the school environment and contributes to the reduction of sedentary behavior, although further technical refinement is still required.**

Keywords-technology; health; mobile app.

I. Introduction

The socio-technological context has strongly influenced the population's lifestyle in recent decades. In the contemporary period, marked by the predominance of the digital age, it is clear that society has become dependent on mobile devices, resulting in excessive use of screens in everyday life, which, in turn, can contribute to problems, such as reduced physical activity and increased sedentary lifestyle, compromising physical and mental health, especially among young people [1].

Several factors contribute to low adherence to physical activity, including the lack of suitable spaces, prolonged technology use, and inadequate public policies to encourage healthy habits [2]. Similarly, the COVID-19 pandemic has worsened this situation by restricting movement in collective spaces, intensifying social isolation, and increasing the population's screen time [3].

Specifically among school-age adolescents, the rates of physical inactivity are worrying. The World Health Organization (WHO) studies indicate that four out of five young people aged 11 to 17 are considered insufficiently active [4]. This scenario is exacerbated by the combination of factors, including intense school workloads, the lack of incentives from educational institutions, and excessive mobile device use [5]. Given this, it is necessary to invest in actions that promote regular physical activity among these young people to improve both their physical and mental health and consequently enhance their academic performance [6].

However, the use of mobile devices in everyday life should not be considered an entirely negative trend, as these resources can help encourage physical activity, depending on how they are used. Studies show that health (or mobile health) applications can be effective in promoting healthier habits, especially among physically inactive individuals [7]. Internet-based technologies, including body movement sensors, can be considered a promising, scalable approach to addressing the high prevalence of physical inactivity [8].

Given this scenario, it is essential to find solutions that mitigate the impacts caused by physical inactivity. Adequate physical exercise improves the health of the body, which consequently benefits brain health, increasing mental well-being and reducing symptoms of anxiety and depression. In this context, we present the "Mobility" application, developed to encourage physical activity among high school students, along with the findings of its initial analysis regarding the effectiveness of this proposal.

This paper is organized as follows: Section II presents related work; Section III describes the application and its functionalities; Section IV presents the results of the experimental evaluation; Section V discusses the results achieved through the evaluation; and Section VI brings the conclusion and future directions.

II. Related work

It is essential to find solutions that mitigate the harm caused by a sedentary lifestyle to health, making it necessary to develop approaches that promote health promotion, especially for the younger population. To consolidate this study, articles investigating the relationship between mobile applications and health promotion were reviewed in the literature.

The study by Zhang et al. [7] conducted a systematic review of the effectiveness of mobile applications (mHealth) in promoting physical activity and reducing sedentary behavior among physically inactive individuals in the health area. The

 ISBN: 978-1-68558-312-5

research analyzed nine clinical studies involving a total of 1,495 participants, comparing application-based interventions with broader approaches that combined the use of apps with other complementary strategies, such as educational sessions, pedometers, or online support groups. The results indicated that the mHealth interventions significantly increased daily physical activity time (an average of 8.72 minutes/day) and reduced sedentary behavior time (an average of 90.94 minutes/day).

Similarly, a systematic review and meta-analysis was conducted on the effectiveness of interventions based on digital technologies Stephenson et al. [9], such as computers, mobile devices, and wearables, in reducing sedentary behavior in healthy adults. The results indicated an average reduction of 41 minutes per day in sedentary time between the intervention groups, with a significant impact in the short term (up to 3 months) and a decrease over time. The most commonly used behavior change techniques were reminders and warnings, self-monitoring, social support, and goal setting.

The study by Ueno et al. [10] presents a scoping review to map the effects of instructions based exclusively on mobile health apps (mHealth apps) aimed at reducing sedentary behavior in adults. The results indicated that most apps contributed to a reduction in sedentary time, with decreases observed in activities such as watching television and total sitting time, in addition to an increase in the number of breaks taken throughout the day. The apps analyzed incorporated features such as reminders, visual feedback, goal setting, and motivational messages, although none of them were commercially available.

Recent studies on mobile health monitoring have highlighted several technical challenges, including variability in sensor accuracy, difficulties in integrating with data collection platforms, and malfunctions in real-world environments. The study carried out by Oliveira et al. [11] reports the challenges faced during longitudinal health monitoring with mobile devices in the wild. Among the main lessons learned, the difficulties with participant engagement and the technical limitations of the sensors stand out. These aspects also impact the use of applications designed to promote physical activity in educational environments, such as Mobility.

The studies presented reinforce the need for solutions such as Mobility, which combines technology and gamification to overcome problems related to physical inactivity. Thus, this study stands out for presenting the use of mHealth in the school context through the integration of gamified resources and a ranking system. Self-monitoring, feedback, rewards, as well as goal setting and review, appear in the literature as some of the most important strategies [12]. Unlike other strategies aimed at promoting physical activity among young people, Mobility stands out by integrating technology and gamification, fostering social interaction among students through healthy competition. It also adapts seamlessly to the school environment, enabling teachers to monitor students' performance in real time through challenge rankings.

III. Mobility App

It is worth noting that this study builds upon a previously published study on the Mobility application, presenting enhancements to the tool's functionalities and new experimental data [13]. Furthermore, the project has been submitted to the Research Ethics Committee (REC) and is currently under review.

Considering the challenges identified about physical inactivity among adolescents, the Mobility application was developed to encourage healthier habits within the school environment. The following are the characteristics of the application, its operation, and the principles that guided its development as a tool to encourage physical activity practice.

The Mobility application was designed with a primary focus on the 17 UN Sustainable Development Goals (SDGs), aiming to align with SDG 3 (Health and Well-being): "Ensure healthy lives and promote well-being for all at all ages" [14].

The application was developed on the Kodular platform, which enables the creation of Android applications through block programming. This approach facilitates the creation of software, especially in educational contexts, as seen with Mobility, which high school students developed.

Firebase Realtime Database was used to store data related to physical challenges and registration/login in the app. Chosen for its ability to provide data in real-time, as well as allow easy integration into Kodular, allowing ranking and challenge information to be updated instantly.

Mobility was designed to integrate seamlessly into school routines, leveraging available resources and capitalizing on young people's familiarity with mobile devices. Its interface is intuitive and seeks to motivate users through its color palette. Orange is associated with energy and vitality, while blue is often used to convey strength and security.

A. Features

The application has the following functions:

- **Registration/Login** (Figure 1): To access the Mobility application, students must complete an initial registration, providing their name, email, password, age, and gender. After this step, access to the platform is done by logging in with the previously registered email and password credentials.

Figure 1. Home screen, registration and login.

- **BMI Calculator** (Figure 2): Allows students to calculate their Body Mass Index based on weight and height data, checking if they are at the correct weight.
- **Physical Challenges** (Figure 2): There are four tasks available on Mobility, which encourage the practice of physical activity in a fun way within the school structure.

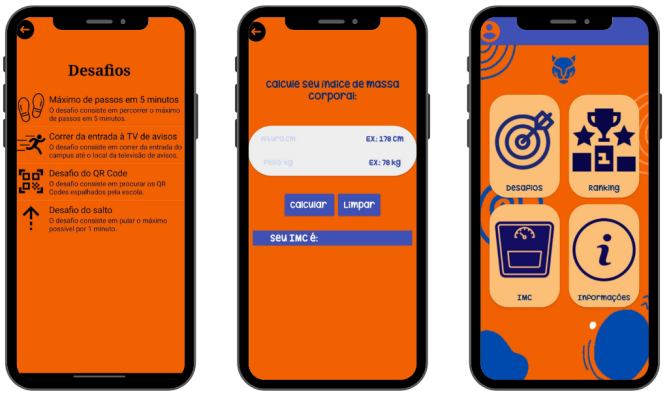


Figure 2. BMI Calculator, Challenges Screen, Features Screen.

- **Ranking:** Allows students to access their position in the ranking of the challenge completed. This system aims to encourage engagement in the application through healthy competition while also facilitating a reward system based on points.
- **About:** Area dedicated to talking about Mobility and its founders.

B. Challenge Modalities

The application has the following challenge modalities:

- **Maximum steps in 5 minutes** (Figure 3): Consists of walking as much as possible for five minutes. The participant with the highest score takes the lead in the ranking.
- **Hallway to TV Announcement Race** (Figure 3): Students must walk the route from the school entrance to a fixed point in the shortest possible time.
- **QR Code Challenge** (Figure 4): Simulating a treasure hunt, students must look for QR Codes spread throughout the school environment. Each code directs to a question, and for each correct answer, the student earns 10 points. The winner is the one who accumulates the highest score.
- **Jump Challenge:** (Figure 4): Proposes that the student performs the most significant number of jumps possible within one minute. The student with the most repetitions wins the challenge.

The development of Mobility demonstrates that it is possible to combine technology and health in an accessible and effective solution in the school context. The following section presents the methods used to conduct the preliminary evaluation of the application, along with a discussion of the results obtained.

IV. Evaluation

To verify the usability and functionality of the application, an evaluation was conducted with 23 first-year high school students, examining the functions available within the application.

Figure 3. Challenge screen: Maximum steps in 5 minutes, Hallway to TV Announcement Race.

Figure 4. Challenge screen: QR Code Challenge, Jump Challenge.

This stage aimed to identify the level of engagement of students with the application, as well as its practical functioning.

The assessment was based on the ISO/IEC 25010 standard [15], following its five steps, as proposed in previous studies on software quality evaluation [16]:

- **Activity 1: Define the assessment**

At this stage, the purpose of the evaluation was established: to analyze the suitability of Mobility in the school context.

- **Activity 2: Design the assessment**

Criteria were defined based on the quality characteristics provided for in ISO/IEC 25010:

Usability: ease of use and understanding of the interface;
Efficiency: time required to complete tasks;
User satisfaction: general opinion about the tool;
Reliability: stability of the application during testing;
Functionality: suitability of functions in relation to the

proposal.

To validate this stage, a form created through Google Forms was designated at the end of the test, containing six questions, with the first four questions allowing answers of 'yes', 'partially', or 'no'. The remaining questions included the participant's written opinion about the application. At the end, they were asked to evaluate the application with a score ranging from 1 to 5, with 1 (does not meet the programmed criteria) and 5 (is suitable for the school context). The questions in the questionnaire are:

1) Does the software perform all the expected functions?
2) Does the application frequently fail?
3) Is the interface easy to navigate?
4) Can the software be used with the infrastructure available at the school?
5) Mobility's strengths:
6) Points for improvement:

- **Activity 3: Plan the assessment**

It was decided that the assessment would take place at the educational institution itself, requiring the use of an Android device.

- **Activity 4: Carry out the assessment**

During this stage, students registered on the platform and, upon logging into the app, explored the available resources. They used the BMI calculator and participated in the activities proposed by Mobility. The test was conducted over two days in different weeks, allowing the class's performance to be assessed more accurately. At the end of the assessment, participants were asked to complete the questionnaire in Activity 2.

- **Activity 5: Complete the assessment**

The students who participated in the evaluation chose to participate voluntarily, having been informed about how Mobility works, including its proposal and the evaluation method.

V. Results and Discussion

The students who participated in the evaluation did so voluntarily and were informed about the functioning of Mobility, including its purpose and the evaluation method.

The challenges were assessed in terms of their functionality during the evaluation and the level of student interest. Regarding student engagement during the activities, a high level of engagement in the activities was observed. During the first challenge (5-minute step challenge), participants spread out throughout the school and competed healthily to take first place in the app's ranking.

Subsequently, the QR Code Challenge was implemented, distinguishing itself as the longest-lasting activity among the proposed challenges. This stage was notable for its integration of technology and physical movement, as each code was linked to a unique Google Sheets URL. Students spontaneously organized themselves into small groups and began searching for the codes distributed across various locations within the school. This phase not only facilitated active physical movement throughout the school environment but also fostered interaction among participants, promoting cooperation, socialization, and sustained engagement in physical activity throughout the task.

At the end of the challenge, it was observed that the students exhibited signs of physical fatigue due to the intense physical activity required by the previous tasks. Therefore, a short break was granted to allow participants to rest before continuing with the next challenge, respecting the limits of the participants and preserving the quality of the evaluated experience.

After the break, the next challenge to be completed was the "maximum jumps in one minute" challenge. This activity was performed individually, but during the execution, the students' engagement was noticeable, even though some showed signs of fatigue.

Finally, the last challenge proposed consisted of running from the school entrance to the announcement TV, recording the shortest possible time. However, this activity had technical limitations, as the geolocation system (GPS) did not work correctly on all devices. Only one of the cell phones used was able to record the route correctly.

The Global Positioning System (GPS) operates through a network of satellites, providing the geographical location of any point on Earth via latitude and longitude coordinates. However, GPS signals generally do not perform accurately in indoor environments (i.e., covered spaces such as houses or buildings) [17]. When used within the school environment, the sensor did not function properly, thereby limiting the users' experience.

Furthermore, since the application was developed on the Kodular platform—which generates only APK files—it is compatible exclusively with the Android operating system. As Kodular does not support the generation of executable files for iOS, the application could not be made available on the App Store or installed on Apple devices. Consequently, students using iOS devices faced access restrictions and had to rely on borrowed Android devices.

Mobility was designed as a secure digital environment that ensured the integrity of the proposed activities. Each QR Code used within the application is linked to a unique URL hosted in a Google Sheets document, which prevents tampering or modification. Therefore, each code corresponds to a specific question, and any attempt to alter the link renders the code unreadable by the application. For example: Code 1 $\rightarrow$ Link 1 $\rightarrow$ Question 1.

Despite some technical limitations, such as sensor failures on specific devices, which especially impacted the GPS-based challenge, the execution of the tasks went well, demonstrating the potential of the Mobility app as a tool to encourage physical activity in the school environment. Furthermore, because it was developed on the Kodular platform, the app is currently only compatible with the Android system. Therefore, students using iOS faced access restrictions and had to use borrowed devices.

After completing the physical challenges proposed by the Mobility app, participants answered an evaluation questionnaire aimed at identifying their perception of the tool's effectiveness, usability, and functionality. In total, 23 students completed the form.

1) Does the software perform all the functions it is supposed to?

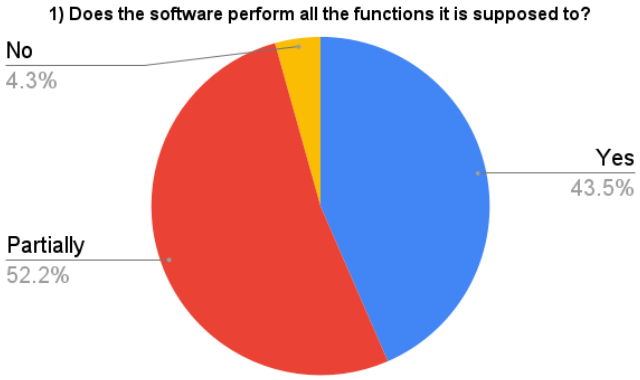


Figure 5. Results of the first question.

The majority of students (52.2%) believe that the application only partially meets the expected functionalities, which can be attributed to failures related to the use of GPS, which limited the experience of the Mobility activities.

2) Does the application crash frequently?

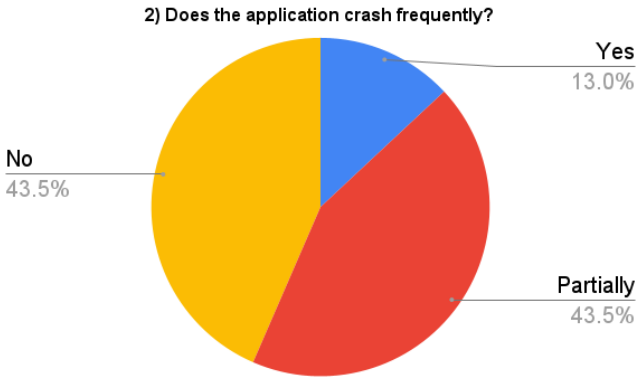


Figure 6. Results of the second question.

For the same reason, there was a balance concerning application failures. While 43.5% stated that there were no frequent failures, the other 43.5% reported partial failures due to the use of GPS. With this result, it means that despite the occasional error, no serious problems were encountered during the evaluation.

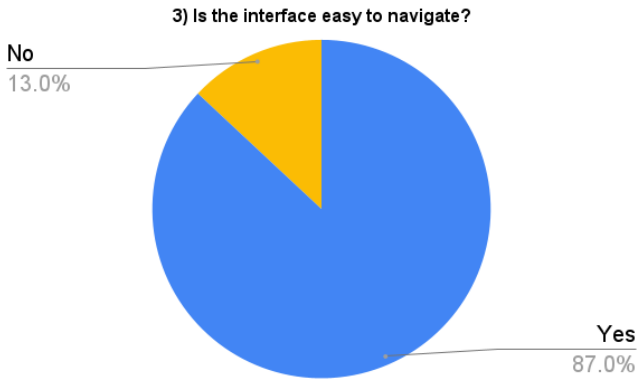


Figure 7. Results of the third question.

3) Is the interface easy to navigate?

On a positive note, 87% of participants found the app's navigation simple and intuitive. This result is a significant feature, as it enables students with varying levels of technology familiarity to interact with the app.

4) Can the software be used with the infrastructure available at the school?

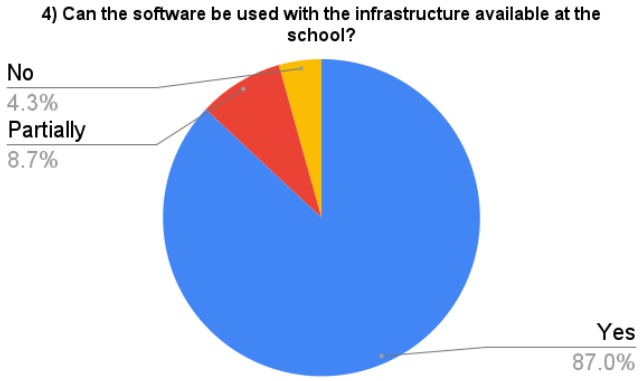


Figure 8. Results of the fourth question.

Regarding compatibility with the school's infrastructure, 87% also stated that the application could be used without significant difficulties, and only a minority reported difficulties: the limitations indicated were related to the quality of the internet connection or the need for Android devices.

Regarding the last two questions, the strengths of Mobility highlighted by the students were the software interface and the activities, which encouraged the practice of physical activities in a fun way. However, they reported the need to create new physical challenges to intensify the practice, in addition to correcting errors related to GPS and making the application available for iOS devices. Finally, Mobility received an average score of 4.22 (on a scale of 1 to 5), indicating a positive acceptance of the tool in the school environment.

VI. Conclusion and Future Work

Based on the results of the Mobility assessment, the persistent objective was to highlight the value of the tool in promoting

health and well-being in the school environment, emphasizing its direct contribution to the UN Sustainable Development Goals (SDGs), particularly those related to health and well-being (SDG 3).

By proposing challenges that encourage physical activity through playful approaches, Mobility stands out as a creative and accessible solution to combat inactivity among young people. The combination of educational technology and gamification promoted student engagement, facilitating learning and participation in the proposed activities.

A. Study limitation

One of the limitations of Mobility is related to technical issues, such as incompatibility with iOS devices, in addition to the limitation of the Running Challenge, which requires the use of GPS.

The evaluation was conducted in a specific setting with only twenty-three students, which does not allow for a general assessment of the application's acceptance among high school students. Another challenge to be addressed is student engagement with the application, as some students may resist using the software due to leading a predominantly sedentary and inactive lifestyle.

B. Future Work

For future research, it is intended to implement Mobility in different schools over a defined testing period, with the aim of analyzing students' reactions and their level of engagement with the application, in order to ensure more objective and reliable results. In addition, strategies will be developed to overcome challenges related to the infrastructure required for the app's use, such as the limited performance of GPS in indoor environments.

REFERENCES

[1] A. Lepp and J. E. Barkley, "Cell phone use predicts being an "active couch potato": Results from a cross-sectional survey of sufficiently active college students", *Digital Health*, vol. 5, p. 2 055 207 619 844 870, 2019.

[2] R. W. Ferreira et al., "Access to public physical activity programs in brazil: National health survey, 2013", *Cadernos de Saude Publica*, vol. 35, e00008618, 2019.

[3] A. C. Becchi et al., "Encouraging physical activity: Nasf strategies amid the covid-19 pandemic", *APS EM REVISTA*, vol. 3, no. 3, pp. 176–181, 2021.

[4] W. H. Organization, *Physical activity*, [Online; accessed 2025-05-12].

[5] K. L. Morton, A. J. Atkin, K. Corder, M. Suhrcke, and E. M. F. van Sluijs, "The school environment and adolescent physical activity and sedentary behaviour: A mixed-studies systematic review", *Obesity Reviews*, vol. 17, no. 2, pp. 142–158, 2016.

[6] M. Monserrat-Hernández, J. C. Checa-Olmos, Á. Arjona-Garrido, R. López-Liria, and P. Rocamora-Pérez, "Academic stress in university students: The role of physical exercise and nutrition", in *Healthcare*, MDPI, vol. 11, 2023, p. 2401.

[7] M. Zhang, W. Wang, M. Li, H. Sheng, and Y. Zhai, "Efficacy of mobile health applications to improve physical activity and sedentary behavior: A systematic review and meta-analysis for physically inactive individuals", *International Journal of Environmental Research and Public Health*, vol. 19, no. 8, p. 4905, 2022.

[8] M. V. L. d. Oliveira, "Mhealth: Possibilities in the field of physical activity and health outcomes", Dissertação (Mestrado em Aspectos Biopsicossociais do Exercício Físico e Aspectos Biopsicossociais do Esporte), Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2018, p. 102.

[9] A. Stephenson, S. M. McDonough, M. H. Murphy, C. D. Nugent, and J. L. Mair, "Using computer, mobile and wearable technology enhanced interventions to reduce sedentary behaviour: A systematic review and meta-analysis", *International Journal of Behavioral Nutrition and Physical Activity*, vol. 14, pp. 1–17, 2017.

[10] D. T. Ueno et al., "Mobile health apps to reduce sedentary behavior: A scoping review", *Health Promotion International*, vol. 37, no. 2, daab124, 2022.

[11] P. A. M. Oliveira, R. M. C. Andrade, and P. A. S. Neto, "Lessons learned from mhealth monitoring in the wild", in *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies, Lisbon, Portugal*, 2023, pp. 155–166.

[12] S. Michie, C. Abraham, C. Whittington, J. McAteer, and S. Gupta, "Effective techniques in healthy eating and physical activity interventions: A meta-regression", *Health Psychology*, vol. 28, no. 6, pp. 690–701, 2009.

[13] M. L. Moraes et al., "Mobility: Promoting health and physical activity in school environments",

[14] United Nations, *The 17 goals | sustainable development*, [Online; accessed 2025-05-25].

[15] I. 25000, *Iso/iec 25040*, [Online; accessed 2025-05-25].

[16] O. Gordieiev, V. Kharchenko, N. Fominykh, and V. Sklyar, "Evolution of software quality models in context of the standard iso 25010", in *Proceedings of the Ninth International Conference on Dependability and Complex Systems DepCoS-RELCOMEX*, Brunów, Poland: Springer International Publishing, Jun. 2014, pp. 187–196.

[17] G. Dedes and A. G. Dempster, "Indoor gps positioning-challenges and opportunities", in *VTC-2005-Fall. 2005 IEEE 62nd Vehicular Technology Conference, 2005.*, IEEE, vol. 1, 2005, pp. 412–415.

Measurability: Toward Proactive Scalable Cybersecurity Management of Large National Infrastructure – USA Healthcare

William Yurcik†
Centers for Medicare & Medicaid Services (CMS)
Baltimore, MD USA
william.yurcik@cms.hhs.gov

Stephen North
Infovisible
Oldwick, NJ USA
scnorth@gmail.com

Rhonda O'Kane
BitSight Technologies
Boston, MA USA
rhonda.okane@bitsighttech.com

O. Sami Saydjari
Dartmouth College
Hanover, NH USA
sami.saydjari@dartmouth.edu

Fabio Roberto de Miranda
Rodolfo da Silva Avelino
Insper Institute of Education and Research
São Paulo, Brazil
{fabiomiranda, rodolfosa1}@insper.edu.br

Gregory Pluta
University of Illinois
Urbana-Champaign, USA
gpluta@illinois.edu

***Abstract*— The current state of cybersecurity protection is reactive response to solving problems as they arise. Many efforts have been undertaken to raise cybersecurity protection awareness and legal liability in an effort to reduce the number and impact of problems, however, problems continue to arise and the result of these improvement efforts is unmeasured and unknown. We propose a new paradigm where cybersecurity posture can be proactively baselined (on a large scale) and then strategic interventions to improve cybersecurity posture can be measured with quantitative results (on a large scale). To demonstrate, we focus on USA healthcare which is currently estimated to be about 17% of the U.S. economy. We show the cybersecurity posture of a large critical national infrastructure can be quantitatively baselined. We accomplish this with an implementation combining the use of data reducing ratings and data visualization techniques. To our knowledge this new paradigm results in the first Internet security management findings for a large national infrastructure.**

Keywords: critical infrastructures protection; cybersecurity quantification; cybersecurity management; hospital cybersecurity.

I. INTRODUCTION

Cybersecurity management encompasses all the planning, implementation, operations, incident response, and remediation required to protect networked resources and ultimately data within an enterprise. Cybersecurity management techniques vary based on the unique enterprise environment and the skills and experience of the people responsible for it.

One powerful management technique that can be employed in the security management domain is the use of quantitative measurement to provide mathematical analysis that are objective, replicable, and enable meaningful precise comparisons [1]. Two influential management theorists, Peter Drucker and W. Edwards Deming, have been falsely attributed with the phrase "If you can't measure it then you can't manage it". This misattribution is understandable since it mirrors both their work. Demings is recognized as the father of total quality management based on continuous measured improvement [2].

However, the use of quantitative measurement for security management is fundamentally challenging for the issues also illuminated by Drucker and Demings- What is important to be managed? What can be measured? Are measurements available for things important to be managed? Can measurements be created for important things to be managed that are currently unmeasured? Can we measure efficiently? We want to measure things important to be managed, not just where measurements are available. What gets measured may get managed even if what we want to manage is not always measurable. Drucker commented directly on this dilemma – "*What gets measured gets managed – even when it's pointless to measure and manage it, and even if it harms the purpose of the organization to do so"* [3].

The current state of cybersecurity management is reactive solving problems as they arise. Cybersecurity & Infrastructure Security Agency (CISA) security management mandated for U.S. Federal agencies consists of enterprise dashboards for critical infrastructures showing system vulnerabilities that have been identified but unpatched and/or otherwise not yet remediated [4]. Log-based security management (e.g., Splunk) and SIEM-based security management (Security Information & Event Management e.g., product RSA NetWitness) consist of enterprise dashboards of prioritized alarms. Compliance-based security management (e.g., Federal Information Security

† Corresponding Author; Official Organizational Disclaimer: "The views presented herein do not represent the views of the Federal Government."

Modernization Act FISMA controls) use an audit control checklist in comparison with a security standard (e.g. NIST 800-53), however, audit controls are not weighted such that one documentation finding is the same as one unimplemented technical control finding leading to the characterization of "check-the-box". Lastly, outsourcing security management to an external entity only transfers responsibility to contractual agreements.

Drucker did actually state, "*The best way to predict the future is to create it"* [3]. In the case of security management, predicting the future is *proactively* creating resilience against future unknown cyberattacks - as opposed to focusing entirely on reactively remediating past known cyberattacks.

We propose a new security management paradigm where cybersecurity posture can be proactively baselined (on a large scale) and then strategic interventions to improve cybersecurity posture can be measured with quantitative results (on a large scale).

To further unpack scalability at a large scale, even if able to produce quantitative security measurements, and given automation support, the volume of security metric information at some point will become too large for human decision-making to take into account relationships, interactions, and emergent properties when making strategic security decisions.

There are two general techniques that can be leveraged to help address scalability. First, numerical data reduction techniques can combine multiple data measurements from multiple sources while retaining underlying information. Second, humans have extraordinary visual processing capabilities, especially for pattern recognition changes, capabilities estimated to be about 10 Mbps with brain reaction times on the order of 150ms [5] [6].

In order to achieve scalable security management, we converged on a two-stage approach consisting of (1) numerical data reduction techniques to reduce data volume and (2) data visualization techniques designed to present information to human decision-makers. After initial proof-of-concept experiments and in-house trial-and-error adjustments, we implemented this two-stage approach for a complex real-world environment.

The remainder of this paper is structured as follows. In Section II, we describe how cybersecurity ratings are derived from empirical security metric measurements. In Section III, we use cybersecurity ratings to baseline large and defined U.S. hospital systems. We end with a summary in Section IV.

II. CYBERSECURITY RATINGS

Cybersecurity ratings based on security metrics can be viewed as a numerical data reduction technique for security metrics, directly analogous to how a credit score is used to encompass overall credit risk by a creditor, and similar to how the current price of a stock or bond encompasses corporate financial reports and market conditions [7].

BitSight invented the ratings industry by creating a transparent algorithm based on security metrics to produce quantitative security scores (ranging from 200-900) for systems/organizations. BitSight is unique in that it incorporates large-scale analysis based on Internet traffic gathered outside of an organization's security perimeter (not egress/ingress traffic) in addition to low frequency network and port scans and open source information.

The previous intuitive analogies we used for cybersecurity rating scores have become physically manifest in the real-world when one of the two largest financial credit rating companies in the world (Moody's) bought an equity stake in BitSight. On 9/13/2021 Businesswire announced Moody's Corporation (New York Stock Exchange NYSE symbol: MCO) invested $250M in BitSight and BitSight acquired VisibleRisk, a cyber risk ratings joint venture created by Moody's and Team8, a global venture group.

Figure 1 shows the security metrics and corresponding weights BitSight uses to calculate their ratings. BitSight groups these security metrics (aka risk vectors) into four categories: (1) Diligence, (2) Compromised Systems, (3) User Behavior, and (4) Public Disclosures. The largest weight is the Diligence risk vector (70.5%) which measures 11 different metrics for best practice implementation. The 4 additional metrics listed under Diligence are currently in beta and do not affect ratings. The next largest weight is the Compromised Systems risk vector (27%) which measures 5 different metrics for evidence of preventing (or lacking to prevent) malicious or unwanted software. The smallest weight is the User Behavior (2.5%) risk vector which measures 3 different activity metrics (open ports, password re-use, and file sharing traffic). Unlike the other three risk vectors, the absence of a Public Disclosure in open source reports does not positively boost ratings but the report of compromise or breach will have a negative impact on ratings.

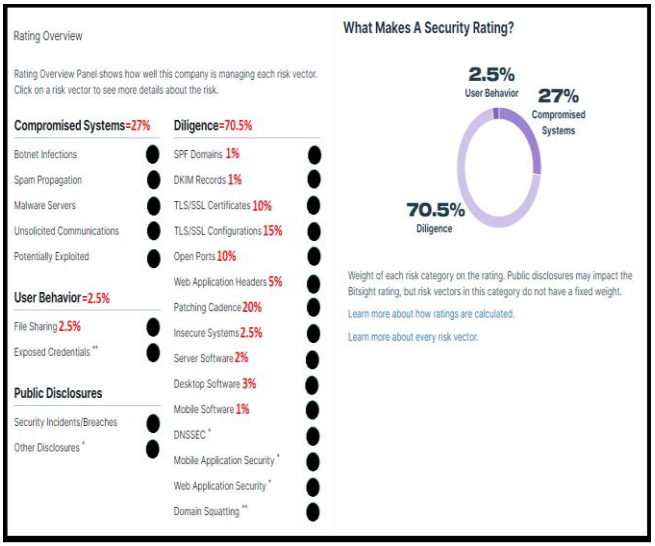


Figure 1. BitSight 2023 Rating Algorithm (used with permission).

For trust and transparency, BitSight publishes its ratings algorithm and annually makes revisions (security metrics and corresponding weights) given user input, changes in the Internet threat environment, and security metric measurement improvements. This follows the well-established model used by other ratings organizations in securities and insurance.

As significant as it is to incorporate an overall cybersecurity risk assessment into one number, a BitSight rating is still only a single data point in time. For human decision-making it is often more important to know where a rating is trending in time as opposed to where it currently stands at the moment. BitSight provides ratings trend sparklines for a one year time period.

Figure 2 is an example BitSight rating trend sparkline annotated with notes documenting rating inflection points. The shaded horizontal rectangle is the expected ratings range where organizations of the same type should be operating. Trends over time are the dominant metric in all ratings organizations especially securities, credit, and insurance. In fact, the Wall Street Journal publishes not only stock prices but individual stock sparklines as demanded by their customers so (as the adage goes) investors and speculators desire to "buy low and sell high".

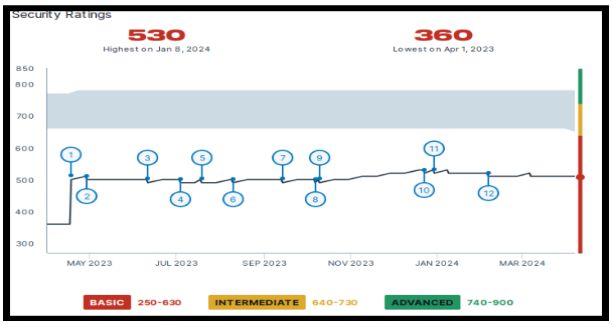


Figure 2. BitSight Annotated Sparkline (used with BitSight permission).

III. BASELINING CYBERSECURITY OF USA HEALTHCARE

At this point, we will pivot to demonstrate how ratings can be used to perform security management on actual infrastructures larger in size than previously possible. Out of possible application domains, we have selected to assess the overall security posture of the USA healthcare sector.

Healthcare includes all organizations, people, and actions whose primary intent is to promote, restore, and/or maintain health. This includes medical providers (doctors/dentists/mental-health-professionals), out-patient urgent care, community clinics, nursing homes, specialized medical equipment providers, health insurers, the pharmaceutical industry, and different types of hospitals.

USA healthcare covers a current population of 333M people, with private group insurance plans covering about 66% of the population, Medicaid covering 89M, Medicare covering 64.5M, the Affordable Care Act covering 21M, and 26M people with no health insurance [8]. As of May 2022 exactly 64,553,288 people were enrolled in Medicare and exactly 88,978,791 people were enrolled in Medicaid and Children's Health Insurance Program (CHIP) [8]. About 12M individuals are dually eligible for both Medicare and Medicaid, so are counted in the enrollment figures for both programs [8]. In January 2024 the Affordable Care Act's Health Insurance Marketplace reached 21M for the 2024 plan year [8]. In September 2023, the U.S. Census reported that for 2022 the number of uninsured U.S. citizens reached a record low of 26M or 7.9% [8]. Note that, due to significant overlaps in coverage, these numbers do not add to the current USA population for the year of study [8]. In 2022, USA healthcare expenditure accounted for $4.5 trillion which is 17.3% of the U.S. Gross Domestic Product (GDP) [8].

To tangibly assess the security posture of USA healthcare, we converged on hospitals as the central point touching every part of the industry – most providers have hospital privileges and hospitals are typically the parent organization of subsidiary activity such as associated out-patient services/facilities. We used multiple open source authorities to assemble a database of 7,490 USA hospitals hosted at the University of Illinois which has been vetted multiple times. Figure 3 shows all USA hospitals mapped to their geographical coordinates in the continental USA.

According to the American Hospital Association, a hospital is state-licensed institution whose function is to provide diagnostic and therapeutic patient services for medical conditions, with organized physician staff and registered nurses. The functional hospitals we are tracking include general hospitals, Short-Term Acute Care Hospitals (STACH), Long-Term Acute Care Hospitals (LTACH), Inpatient Rehabilitation Facilities (IRF), Skilled Nursing

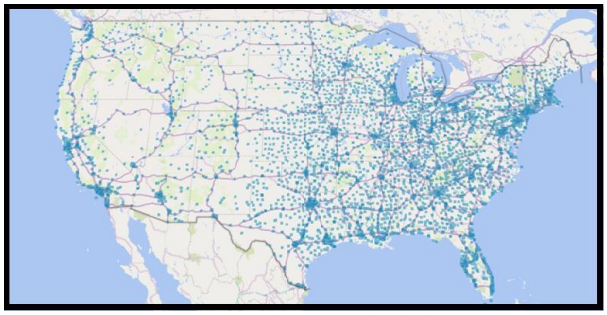


Figure 3. USA Hospitals Geographical Placement.

Facilities (SNF), short stay hospitals, behavioral hospitals, psychiatric care hospitals, children's hospitals, women's hospitals, teaching hospitals, and specialty care hospitals (cancer care, eye surgery, etc). Formal categories of hospitals include Acute Care/Critical Access Hospitals (ACH, fewer than 25 in-patient beds and greater than 35 miles from the next nearest hospital) and Safety-Net Hospitals (designated by the proportion of charity care provided). In addition to leveraging authoritative sources, we identified and vetted hospitals based on healthcare facilities containing in-patient beds, the word "hospital" in their title (which is regulated by state authorities), and Internet website presence.

We subdivided USA hospitals into five separate systems for analysis: (1) Indian Health Service Hospitals, (2) Veterans Health Administration Hospitals, (3) Defense Health Agency Hospitals, (4) Interstate Hospital Systems, and (5) Intrastate Hospital Systems. These five hospital systems include 69% of all the hospitals in the USA, with the remaining hospitals being independent unaffiliated hospitals.

A. Baseline – Indian Health Service (IHS)

IHS is the primary Federal healthcare provider (administered by the U.S. Department of Health & Human Services) for Federally-recognized American Indian tribes and Alaskan natives consisting of approximately 2.6 million people belonging to 574 tribes in 37 states. In the role of primary healthcare provider, IHS provides a comprehensive health service delivery system consisting of 24 IHS hospitals and 22 Tribal hospitals; 51 IHS Health Centers and 279 Tribal Health Centers; and 59 Alaska Village Clinics.

From this IHS/Tribal facility mix, we identified and processed 46 in-patient hospital/medical center facilities located in ten different states containing a cumulative total of 1,620 beds. Of the 46 in-patient facilities, five IHS and nine Tribal Hospitals are critical access hospitals, and one of the Tribal Hospitals is an inpatient rehabilitation facility. Figure 4 shows all IHS hospitals mapped to their geographical coordinates in the continental USA. There are 7 IHS hospitals in Alaska not shown in Figure 4.

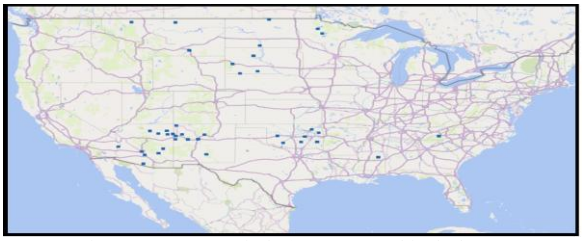


Figure 4. IHS Hospitals (46) Geographical Placement.

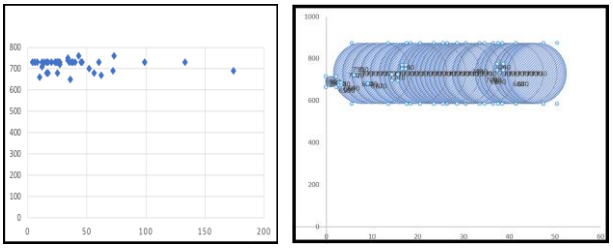


Figure 5. IHS Hospital Ratings (46) vs Hospital Size.

The BitSight rating for each of the 46 in-patient IHS/Tribal facilities are shown in Figure 5 as a function of hospital size. Rightmost Figure 5 is a representation of the number of assets (URLs, IP addresses, domain names) being monitored at each IHS hospital – the more assets the larger the dot/circle. Figure 6 breaks out BitSight ratings and hospital sizes in separate histograms.

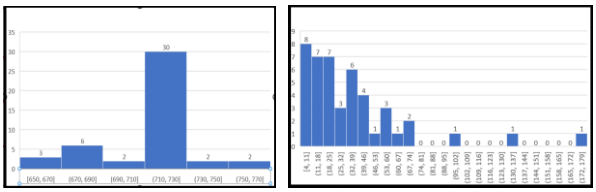


Figure 6. IHS Hospitals (46) Ratings vs Hospital Size.

Leftmost Figure 6 frequency distribution shows the IHS hospital rating scores bundled into histogram bins sizes of 20. The vertical axis is frequency. The IHS hospital system mean rating score is 719.78 with scores ranging from 650-760 (110 range), median/mode of 730, a negative skew of -1.23 (median/mode higher than mean) with more scores higher than the mean, and a 95% confidence interval around the mean of 712.53 - 727.03. Twelve IHS hospitals fall outside-below the mean 95% confidence interval.

The rightmost Figure 6 histogram shows the distribution of IHS hospital sizes as measured by in-patient beds in bins sizes of 7 beds. While the mean size of an IHS hospital is 36 in-patient beds, almost half of the IHS hospitals are smaller critical access hospitals defined as being less than or than or equal to 25 in-patient beds.

B. Baseline - U.S. Veterans Health Administration (VHA)

VHA is the largest healthcare system in the world providing healthcare for about 9 million non-active/discharged veterans of the U.S. military annually at 1,321 healthcare facilities, including 172 medical centers, 1,138 community-based outpatient clinics, and 134 Community Living Centers (e.g. nursing homes) [9]. All VHA healthcare facilities are owned and operated by the U.S. Department of Veteran Affairs and the approximate 350,000 VHA healthcare staff are Federal employees making them the second largest workforce cohort in the U.S. government [9] [10]. Multiple reports show VHA hospitals provide quality healthcare that is equal to, and often better than, healthcare provided by private sector hospitals [11] [12].

Figure 7. VHA Hospitals (168) Geographical Placement.

We processed 168 in-patient VHA hospital/medical center facilities located in 51 states (including Washington D.C.) containing a cumulative total of 38,296 beds. Figure 7 shows all VHA hospitals mapped to their geographical coordinates. Not shown in Figure 7 are VHA hospitals in Alaska(1), Hawaii(1), and Puerto Rico(1).

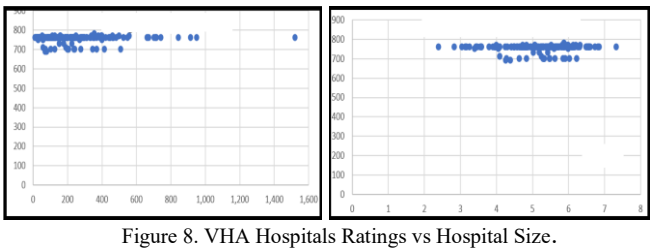


Figure 8. VHA Hospitals Ratings vs Hospital Size.

Figure 8 shows a scatter plot of ratings for VHA hospitals. The vertical axis is ratings and the horizontal axis is the number of in-patient beds within each of the 168 VHA hospitals (leftmost horizontal axis is straight scale, rightmost horizontal axis is scaled log base e). Figure 9 breaks out the

ratings and hospital sizes for VHA hospitals into separate frequency distribution histograms.

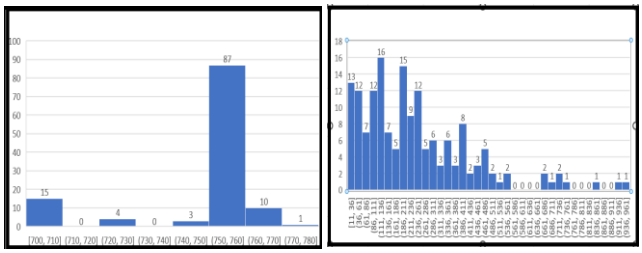


Figure 9. VHA Hospitals (168) Ratings & Hospital Size.

The leftmost Figure 9 frequency distribution shows the VHA hospital ratings bundled into histogram bins sizes of 10. The vertical axis is frequency. The VHA hospital system mean rating score is 753.78 with scores ranging from 690-780 (90 range), median/mode of 760, a negative skew of -2.27 (median/mode higher than mean) with more scores higher than the mean, and a 95% confidence interval around the mean of 750.81 – 756.74. Twenty-five VHA hospitals fall outside-below the 95% confidence interval for the mean.

The rightmost Figure 9 indicates VHA hospital sizes via a frequency distribution histogram of in-patient hospital beds with a bin size equal to 25. The mean size of a VHA hospital is 248.18 in-patient beds. The large Chillicothe VHA Medical Center in Ohio with 1,522 in-patient hospital beds is included in mean in-patient bed calculation but intentionally omitted in Figure 9 display for data visibility.

C. Baseline - U.S. Defense Health Agency (DHA)

DHA is operated by the U.S. Department of Defense as the healthcare provider for active-duty members of the U.S. military with hospitals and clinics worldwide. About 9.4M active-duty members of the U.S. military use DHA hospitals and clinics with TRICARE military health insurance expenditures representing about 8% of the U.S. Department of Defense (DoD) budget [13].

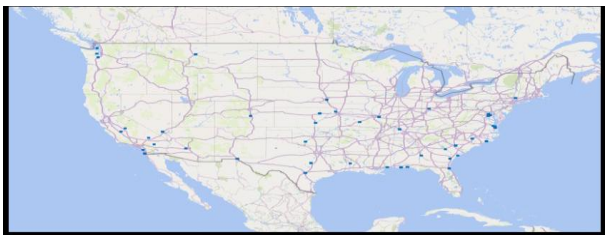


Figure 10. DHA Hospital (48) Geographical Placement.

For the purposes of this paper we will focus only on DHA hospitals located in the USA. We identified and attempted to process 48 in-patient DHA hospital/medical center facilities located in 25 states containing a cumulative total of 8,358 beds. Figure 10 shows all DHA hospitals mapped to their geographical coordinates in the continental USA. Not shown in Figure 10 are DHA hospitals in Alaska(2) and Hawaii(1).

Figure 11 indicates the distribution of DHA hospital sizes with a frequency distribution histogram of in-patient hospital beds with bin size equal to 25. The mean size of a DHA hospital is 181.70 in-patient beds. The large Blanchfield Army Community DHA Hospital at Fort Campbell in Kentucky with 2,100 in-patient hospital beds is included in mean in-patient bed calculation but intentionally omitted from Figure 12 display for data visibility.

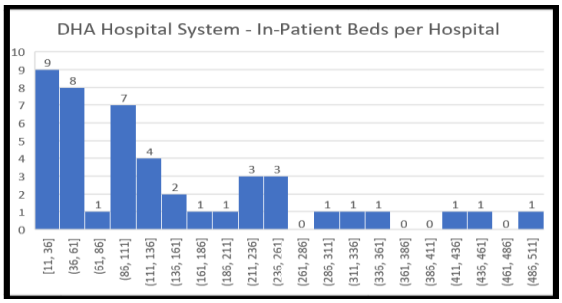


Figure 11. DHA Hospitals (48) – Hospital Size Distribution.

We found the ratings for each of the 48 in-patient DHA Hospitals was pegged at 770 (no variation) and the number of assets detected at each DHA hospital was also pegged at 28 (no variation). DHA facilities are located on secure military installations and all DHA hospitals and clinics are networked together by nine Defense Health Networks which are "dual-hatted" accountable to both DHA and military commands. Given this classified national security environment it is to be expected our attempts to derive ratings were only partially successful with incomplete results.

D. Baseline – Interstate Hospital Systems

USA hospitals are increasingly combining into systems of multiple hospitals sharing the same IT infrastructure – for reasons beyond the scope of this paper. We subdivided these hospitals systems into two categories for analysis: (1) Interstate Hospitals Systems containing hospitals in multiple states and (2) Intrastate Hospital Systems containing hospitals all within one state. This separation based on state boundaries is meaningful since hospital administration is generally governed by state regulations/certifications/laws.

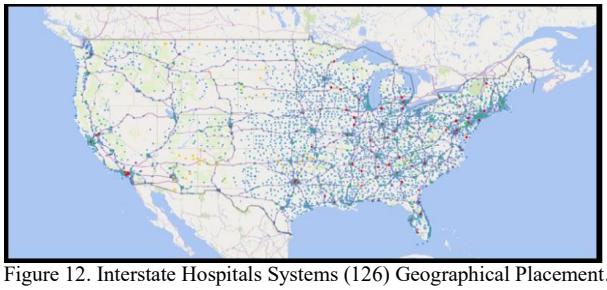


Figure 12. Interstate Hospitals Systems (126) Geographical Placement.

Figure 12 shows the headquarters location of all USA Interstate Hospital Systems mapped to their geographical coordinates in the continental USA. No Interstate Hospital Systems are headquartered in Alaska or Hawaii. We identified 126 Interstate Hospital Systems with a mean size of 21.38 hospitals ranging in size from a two hospital Interstate Hospital System (4 systems) to 84/127/158 hospital Interstate Hospital Systems (HCA Healthcare/Ascension

Healthcare/Encompass Health Interstate Hospital Systems respectively). We identified Interstate Hospital Systems ranging from across only 2 states (60 systems) to Interstate Hospital Systems ranging across 25/37 states (Select Specialty Hospitals/Encompass Health respectively) with the mean number of states in an Interstate Hospital System equal to 4.96 hospitals.

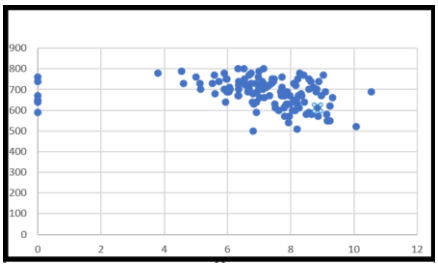


Figure 13. Ratings for Interstate Hospital Systems (126) vs Size.

Figure 13 shows a scatter plot of ratings for USA Interstate Hospital Systems. The rating for each Interstate Hospital System is the combined aggregate score of all hospitals in that system. The vertical axis is ratings and the horizontal axis is the logarithm (base e) of the number of in-patient beds within a hospital.

Figure 14 breaks out ratings and hospital sizes for USA Interstate Hospital Systems into frequency distribution histograms. The leftmost Figure 14 frequency distribution shows the USA Interstate Hospital Systems ratings bundled into histogram bin sizes of 48. The vertical axis is frequency. The USA Interstate Hospital System mean rating is 682.72 with scores ranging from 500 - 800 (300 range), median/mode of 690, a negative skew of -0.52 (median/mode higher than mean) with more scores higher than the mean, and a 95% confidence interval around the mean (684) of 671.00 – 694.72. Fifty USA Interstate Hospital Systems fall outside-below the 95% confidence interval for the mean.

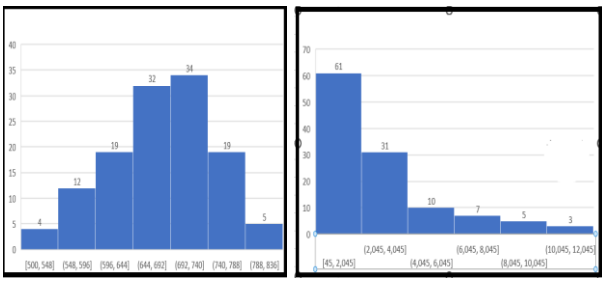


Figure 14. USA Interstate Hospital Systems (126) - Ratings & Hospital Size.

The rightmost Figure 14 indicates USA Interstate Hospital sizes via a frequency distribution histogram of in-patient hospital beds with a bin size equal to 2000. The capacity of in-patient beds within Interstate Hospital Systems range in size from 45 beds (Brightwell Behavioral Health) to 23,557 beds (HCA Healthcare). The mean size of a USA Interstate Hospital System is 3,171.23 in-patient beds, median equal to 1,816 beds and mode equal to 365 beds, with skew equal to 4.76 and stdev equal to 4,598 beds. The large HCA Interstate Health System (23,557 in-patient hospital beds) is included in calculations but intentionally omitted in the Figure 15 display for data visibility.

E. Baseline – IntraState Hospital Systems

With USA state regulations/certifications/laws governing the administration of hospitals, a large number of USA Intrastate Hospital Systems have emerged confined within a single state boundary. We identified 523 Intrastate Hospital Systems across all states ranging in size from two hospitals (167 different Intrastate Hospital systems) to 46 hospitals (Baylor Scott & White Health in Texas) with a mean of 4.92 hospitals. Texas has the largest number of Intrastate Hospitals Systems (41 systems) as well as the most hospitals affiliated within an Intrastate Hospital System (255 hospitals). At the other extreme, Alaska, District of Columbia, and Vermont only have one Intrastate Hospital System, and this one Intrastate Hospital System consists of only one hospital in each of these states.

Figure 15 shows the headquarters location of all USA Intrastate Hospital Systems mapped to their geographical coordinates in the continental USA. Figure 16 shows two scatter plots of BitSight Ratings for USA Intrastate hospital systems. Each Intrastate Hospital System consists of multiple hospitals physically located within the same state and networked together sharing the same IT infrastructure. The rating for each Intrastate Hospital System is the combined aggregate score of all hospitals in that system.

Figure 15. Interstate Hospitals Systems (126) Geographical Placement.

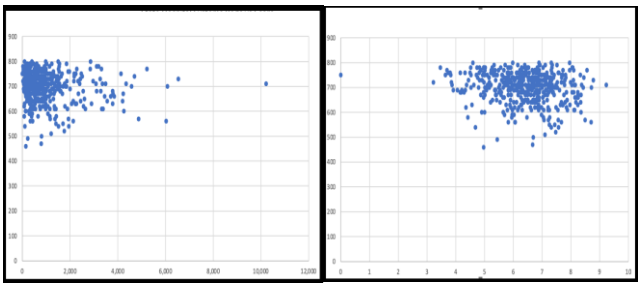


Figure 16. Ratings for Intrastate Systems (523) vs Size.

Figure 16 vertical axis are both ratings, the horizontal axis for the leftmost scatterplot is the number of in-patient beds within a hospital and the horizontal axis for the rightmost scatterplot is the logarithm (base e) of the number of in-patient beds within a hospital. Figure 17 breaks out the

BitSight ratings and hospital sizes for Intrastate Hospital Systems into separate frequency distribution histograms.

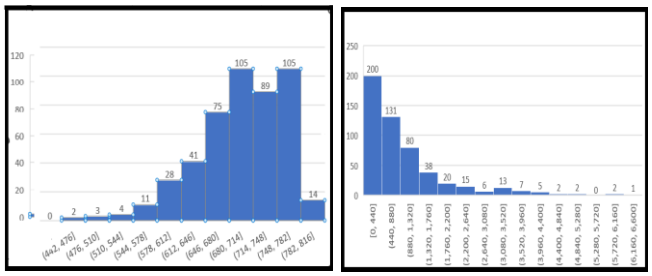


Figure 17. Intrastate Hospital Systems (523) - Ratings and Size.

The leftmost Figure 17 frequency distribution shows the USA Intrastate Hospital Systems security ratings bundled into histogram bins sizes of 34. The vertical axis is frequency. The USA Intrastate Hospital System mean security rating is 699.34 with scores ranging from 460-800 (340 range), median/mode of 710, a negative skew of -0.89 (median/mode higher than mean) with more scores higher than the mean, and a 95% confidence interval around the mean (699.34) of 693.73 – 705.04. Twenty-nine USA Intrastate Hospital Systems fall below the 95% confidence interval for the mean.

The rightmost Figure 17 indicates USA Intrastate Hospital sizes via a frequency distribution histogram of in-patient hospital beds with a bin size equal to 440. The mean size of a USA Intrastate Hospital System as measured in in-patient beds is 955.56 beds (median=626, mode=450, skew=2.95). The capacity for in-patient beds within Intrastate Hospital Systems ranges in size from only 28 beds (Altus Health System in Texas) to 10,214 beds (State of California Health System). New York has the largest number of in-patient beds within an Intrastate Hospital Systems (56,422) while the smallest number of in-patient beds within an Intrastate Hospital Systems is in Vermont (25). The large State of California Hospital System (10,214 in-patient hospital range beds) is included in mean in-patient bed calculation but intentionally omitted in the Figure 17 display for data visibility.

IV. SUMMARY

We have introduced, described, and demonstrated a new cybersecurity rating measurability approach for proactive and scalable security management suitable for infrastructures that are larger in size than previously possible to assess - infrastructures that are national in scale. This new paradigm is based on empirical cybersecurity metric data, and proactive, forward-looking, designed to prevent the next attack rather than focusing on remediating past attacks. For instance, cybersecurity ratings are most sensitive to having time-responsive system patching, and not as sensitive to standardized patching cadences for well-known systems who have regularly been attacked in the past.

Baselining is key to establishing fixed references for measuring progress, managing changes, and assessing performance against schedules and cost. A cybersecurity baseline also provides a reference point for tracking deviations, identifying potential issues, making informed decisions, and ensuring all stakeholders have a unified understanding of goals and expectations.

In this paper we performed proof-of-concept experimental baselining of an actual large national infrastructure (USA hospital systems). Our next step will be to demonstrate how interventions with security investments can be strategically designed to improve security and quantitatively measured for their effectiveness using cybersecurity ratings.

We have also used cybersecurity rating techniques to great effect to investigate other urgent problems. Using cybersecurity ratings again in the USA hospital system context, we discovered three cybersecurity "magnified vulnerabilities" in that a single successful exploit can have an outsized impact on the entire nationwide U.S. healthcare infrastructure [14].

ACKNOWLEDGMENTS

This research was enabled through a cooperative agreement between the University of Illinois at Urbana-Champaign and BitSight. BitSight provided no financial support to this research. Cybersecurity ratings for hospitals presented in this research were processed by BitSight engineers led by Rhonda O'Kane and supported by Tadd Hopkins, Tim Jackson, Tom Linehan, and Will Ricardi. Geocoding was provided by GeoCoder.ca who provided public service access to their geography mapping scripts. Geocoder.ca provided no financial support to this research. Authors Miranda and Avelino were supported by a joint funding support agreement between the Insper Institute of Education & Research and the Computer Science Department at the University of Illinois at Urbana-Champaign.

REFERENCES

[1] National Institute of Standards and Technology (NIST), "Measurement Guide for Information Security: Volume 1 – Identifying and Selecting Measures," NIST SP 800-55, vol. 1. January 17 2024.

[2] M. Best and D. Neuhauser, "W. Edwards Deming: Father of Quality Management, Patient and Composer," Quality and Safety in Health Care, 14(4) Sept 2005. <doi::10.1136/qshc.2005.015289>

[3] P. Drucker, "The Essential Drucker: In One Volume the Best of Sixty Years of Peter Drucker's Essential Writings on Management," Harper Collins Publishers, 2001.

[4] Cybersecurity & Infrastructure Security Agency (CISA), Critical Infrastructure Sectors. retrieved February 9, 2024 from <https://www.cisa.gov/topics/critical-infrastructure-security-and-resilience/critical-infrastructure-sectors>

[5] K. Koch et al., "How Much the Eye Tells the Brain," Current Biology vol. 16, July 25, 2006. <doi:10.1016/j.cub.2006.05.056>

[6] S. Thorpe, D. Fize, and C. Marlot, "Speed of Processing in the Human Visual System," Nature. vol. 381, July 6, 1996. <doi:10.1038/381520a0>

[7] S. J. Choi and M. E. Johnson, "The Relationship Between Cybersecurity Ratings and the Risk of Hospital Data Breaches," J of the Am Medical Informatics Assoc, 2021.

[8] CMS National Health Expenditures (NHE) Fact Sheet. <https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/nhe-fact-sheet#>

[9] Veterans Health Administration (VHA). retrieved March 29, 2024. from <https://www.va.gov/HEALTH/>.

[10] Congressional Budget Office, "Quality Initiatives Undertaken by the Veterans Health Administration," CBO Report, August 2009.

[11] S.M. Asch et al., "Comparison of Quality of Care for Patients in the Veterans Health Administration and Patients in a National Sample," Annals of Internal Medicine, 141(12), 2004. <doi:10.7326/0003-4819-141-12-200413310-00010>

[12] A.N. Trivedi, S. Matula, I. Miake-Lye, P.A. Glassman, P. Shekelle, and S. Asch, "System Review: Comparison of the Quality of Medical Care in Veterans Affairs and Non-Veterans Affairs Settings," Medical Care, 49(1) 2011. <doi:10.1097/mir.0b013e3181f53575>

[13] "Health.mil - The Official Website of the Military Health System," retrieved March 29, 2024 from <https://www.health.mil/About-MHS/OASDHA/Defense-Health-Agency/>

[14] W. Yurcik et al., "Cybersecurity Monitoring/Mapping of USA Healthcare (All Hospitals) – Magnified Vulnerability due to Shared IT Infrastructure, Market Concentration, & Geographical Distribution," ACM CCS Workshop on Cybersecurity in Healthcare (HealthSec), 2024. <doi:10.1145/3689942.3694754>

From Text to Code: Predicting Abbreviated Injury Scale 2015 from Clinical Narratives

Chien-Ming Lee*, Pei-Ling Lee†, Chia-Yeuan Han*,
Joffrey Hsu†, Chuan-Yu Hu†
*Department of Medical Information,
Kaohsiung Medical University Hospital, Kaohsiung Medical University,
Kaohsiung, Taiwan
e-mail: {nomoney.lee | hoganhan2}@gmail.com
†Division of Trauma and Surgical Critical Care, Department of Surgery,
Kaohsiung Medical University Hospital, Kaohsiung Medical University,
Kaohsiung, Taiwan
e-mail: peiling@trauma.tw, {joffrey.hsu | chuanchuan19}@gmail.com

***Abstract*—Accurate coding of traumatic injuries using the Abbreviated Injury Scale (AIS) is very important for trauma registrants. However, manual AIS coding requires trained personnel and is very time-consuming. This study explores the feasibility of using a pre-trained Natural Language Processing (NLP) model to automatically predict complete AIS codes from unstructured diagnostic text entered by emergency physicians. Without additional training or fine-tuning, a publicly available transformer-based model was applied to emergency department narrative data. This preliminary result shows that such models can find clinically relevant information from these free-typing texts and then map to the correct AIS codes. This work highlights the potential of leveraging existing NLP models to assist in injury classification and AIS coding, especially without labeled datasets for training.**

Keywords-abbreviated injury scale; natural language processing.

I. INTRODUCTION

The Abbreviated Injury Scale (AIS) 2015 revision [1] is a globally recognized system for classifying and coding trauma injuries. Accurate and consistent AIS coding is essential for trauma registry maintenance and trauma severity scoring. In particular, the AIS 2015 revision provides detailed seven-digit codes that represent body region, type of anatomical structure, specific nature of injury, level, and severity scoring.

In general, AIS coding is performed manually by trained registrants based on structured clinical data or narrative documentation, including free-text diagnoses from emergency physicians. However, this process is time-consuming, and prone to human error, especially in a busy emergency department setting, the typed narrative diagnoses often vary widely.

Recent advances in Natural Language Processing (NLP), such as the transformer-based language models for biomedical use [2], make it possible to automatically understand complex clinical text narratives. In this study, we demonstrate the feasibility of using a publicly available pre-trained model to automatically predict complete AIS 2015 codes from unstructured diagnostic narratives written by emergency physicians. We focus on zero-shot methods, without labeled data, to evaluate whether such publicly available NLP models can be applied to real-world clinical data.

The remainder of the paper is organized as follows: Section II describes the steps of code mapping procedures. Section III presents preliminary evaluation results. Section IV discusses the findings and limitations. Finally, Section V concludes the paper and future directions.

II. RELATED WORK | METHODS

A. Overview

To predict complete AIS 2015 codes from unstructured emergency department diagnosis narratives, we establish a multi-step matching pipeline that leverages anatomical keyword recognition and NLP models. The approach does not involve model training or fine-tuning. Instead, it uses domain knowledge and semantic similarity scoring to map free-text diagnoses to relevant AIS 2015 codes.

B. Step 1: Extraction of Body Region Keywords

We first find the anatomical keywords mentioned in the AIS coding description. These keywords, such as “skull,” “thorax,” or “femur,” represent anatomical parts involved in trauma and serve as primary factors for matching. The resulting list is manually reviewed to remove ambiguous terms.

C. Step 2: Mapping Body Regions to AIS Codes

For each identified anatomical keyword, we recognize all AIS codes whose descriptions contained that keyword. This generates a mapping table in which each keyword is associated with one or more possible AIS codes. These many-to-many mappings can effectively narrow the comparison range during prediction.

D. Step 3: Body Region Detection in Diagnostic Texts

When processing diagnostic narratives, we scan the entire narrative for the anatomical keywords found in step 1. Depending on the clinical note, zero or more body regions may be detected. These found keywords are used to select candidate AIS codes.

 ISBN: 978-1-68558-312-5

E. Step 4: Semantic Similarity Based Code Selection

For each candidate code retrieved via anatomical keyword mapping, we compute the semantic similarity between the diagnostic text and the code description using a pretrained biomedical language model. Sentence-level embeddings are generated using a transformer-based model trained on clinical and biomedical corpora. Specifically, we use the BioBERT model pre-trained on a large amount of literature in the biomedical domain and is particularly optimized for NLP tasks in the biomedical fields, which is public available on Hugging Face [3]. The AIS code with the highest cosine similarity to the diagnostic text is selected as the predicted result.

III. Results

Table I shows the prediction accuracy of AIS 2015 codes categorized by body region, based on a total of 54 cases. Extremity injuries demonstrates the best performance, with Top-1 accuracy of 70% and Top-5 accuracy of 90%. The Thorax region has the lowest accuracy, without correct Top-1 predictions and only 20% Top-5 accuracy.

Head injuries have no Top-1 hits but reached 50% in Top-5 accuracy. Face/Neck injuries show the moderate accuracy, with 43% Top-1 accuracy and 71% Top-5 accuracy. Abdomen and Spine regions both have relatively low accuracies, around 25-40%. External injuries have better performance, with 60% accuracy for both Top-1 and Top-5. The Other category has a Top-1 accuracy of 25% and a Top-5 accuracy of 63%.

TABLE I. Prediction Accuracy by Body Region

Region	Count	Top-1 Acc.	Top-5 Acc.
Extremity	10	70% (7)	90% (9)
Thorax	5	0% (0)	20% (1)
Head	6	0% (0)	50% (3)
Face/Neck	7	43% (3)	71% (5)
Abdomen	5	40% (2)	40% (2)
Spine	8	25% (2)	25% (2)
External	5	60% (3)	60% (3)
Other	8	25% (2)	63% (5)
All	54	35.2% (19)	55.6% (30)

Overall, the result shows a Top-1 accuracy of 35.2% and a Top-5 accuracy of 55.6% across all body regions, indicating a reasonable performance for predicting AIS codes from free-text emergency department narratives, especially the notable accuracy in extremity injuries.

IV. Discussion | Evaluation

The results indicate that the proposed semantic similarity based approach is feasible for AIS code prediction from free-text Emergency Department (ED) diagnoses, even without the labeled training data. Although the current Top-1 and Top-5 accuracy suggest there is room for improvement, the system still demonstrates potential as a decision support to assist trauma registry personnel in aiding the manual coding process and reducing workload.

Accuracy differs across body regions: better performance for extremity injuries where terminology is more consistent, and lower accuracy for thoracic cases where phrasing is more variable. Current keyword-mapping approach was designed to improve efficiency, but future work may compare against full-text searches with alternative similarity metrics. Preliminary observations indicate that comparison with all AIS code description may result in misclassification for identical injury descriptions and does not significantly improve accuracy; however, further evaluation is needed. In addition, a simple keyword-only baseline was not implemented due to time constraints, it still represents a useful direction for future work to clarify the benefit of semantic similarity.

Practical limitations remain: the approach predicts one AIS code per diagnosis and does not yet support multiple injuries in a single narrative. Small dataset (n = 54) and imbalance body regions also limit generalizability. The reliance on exact keyword matching in body region detection may miss relevant terms due to synonym variation or misspellings. Additionally, since the model is used without any fine-tuning, the results may vary due to different wording by emergency physicians, especially in complex injury scenarios.

V. Conclusion and Future Work

This study proposes a knowledge-driven approach for predicting complete AIS 2015 injury codes from unstructured emergency department diagnostic narratives. By integrating body region keyword detection with semantic similarity scoring using a pretrained biomedical NLP model, the system effectively maps free-text entries to structured AIS codes without relying on labeled training data.

Although the preliminary evaluation result for body region with terminology, especially extremity injuries, shows encouraging Top-1 and Top-5 accuracy, the overall accuracy still has room for improvement. Despite current limitations, such as sensitivity to variable phrasing and the lack of fine-tuning, the results represent the approach's feasibility as a supportive tool within trauma coding workflows.

Future work will focus on expanding the dataset, enhancing body region detection by improving synonym handling and contextual interpretation, and incorporating structured clinical information to improve the accuracy. In the long term, such systems may aid to improve the efficiency, accuracy, and consistency of trauma registry data collection in real-world clinical environment.

References

[1] Association for the Advancement of Automotive Medicine, "The Abbreviated Injury Scale 2015 Revision." Chicaco, IL, USA: AAAM, 2025.

[2] J. Lee et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.
DOI: https://doi.org/10.1093/bioinformatics/btz682

[3] P. Deka, BioBERT-mnli-snli-scinli-scitail-mednli-stsb, Hugging Face, 2021. [Online]. Available: https://huggingface.co/pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb. [retrieved: April, 2025.] Licensed under Creative Commons Attribution-NonCommercial 3.0 (CC BY-NC 3.0). See https://spdx.org/licenses/CC-BY-NC-3.0.

Hypothermia and Its Association with Mortality Among Major Trauma Patients in a Tropical Climate: A Retrospective Study from Southern Taiwan

Pei-Ling Lee*, Chao-Wen Chen†, Chuan-Yu Hu*,
Mei-Yu Pan*, Shu-Fen Ko*, Shu-Chen Mu*
*Division of Trauma and Surgical Critical Care, Department of Surgery,
Kaohsiung Medical University Hospital, Kaohsiung Medical University,
Kaohsiung, Taiwan
e-mail: peiling@trauma.tw, {chuanchuan19 | jadesweet06080527}@gmail.com,
850174@kmuh.org.tw, 920547@ms.kmuh.org.tw
†Department of Emergency Medicine, Faculty of Medicine,
College of Medicine, Kaohsiung Medical University,
Kaohsiung, Taiwan
e-mail: kmutrauma@gmail.com

***Abstract*—Hypothermia is an important factor for poor prognosis after trauma, and it may still occur even in tropical climates. To explore the incidence of hypothermia in patients with severe trauma in southern Taiwan and its association with mortality risk, trauma registration data of a critical emergency hospital in Kaohsiung from 2023 to 2024 were analyzed, and patients with Injury Severity Score (ISS) >15, a recorded emergency department temperature between 32.0°C and 37.9°C, and hospitalization were included. Patients were subsequently categorized into two groups: hypothermia (32.0–34.9°C) and normothermia (35.0–37.9°C). A total of 1324 patients were included, of which 31 (2.34%) had hypothermia, and the mortality rate was 51.6%. Hypothermia was significantly associated with Glasgow Coma Scale (GCS), Intensive Care Unit Length Of Stay (ICU LOS) and death (p<0.005). Hypothermia still occurs in patients with severe trauma in tropical regions, and their mortality risk is significantly increased. Hypothermia should be listed as an important indicator for the initial treatment of trauma, and early intervention can help improve prognosis.**

Keywords-lethal diamond; major trauma; traumatic hypothermia; trauma registry.

I. Introduction

Hypothermia, together with acidosis, coagulopathy, and hypocalcemia, forms the "Lethal Diamond," a cluster of physiological derangements that accelerate trauma mortality. Hypothermia impairs cardiac contractility, induces arrhythmias, disrupts coagulation, and weakens immune response, and has been consistently linked to poor outcomes in trauma patients [1]. It is defined as a core body temperature below 35°C. Literature reports show that up to two-thirds of severely injured adults present with hypothermia upon arrival at the emergency department [2]. However, data from the American College of Surgeons-Trauma Quality Improvement Program (ACS-TQIP) registry indicate a lower incidence (1%) in the United States, though mortality risk increases significantly [3]. Climate appears to influence incidence, with 12.6% in temperate zones [4], 5.7% in subtropical regions such as Australia [1], and even measurable rates in the Middle East [5]. Kaohsiung City, located in southern Taiwan, has a tropical monsoon climate with annual temperatures ranging 15–32°C [6]. This study explores the incidence of hypothermia in trauma patients in a tropical setting and its association with mortality. The remainder of the paper is organized as follows: In Section II, we describe the steps of methods. Section III presents the study results. Section IV discusses the findings. Section V discusses the limitations. Finally, Section VI discusses the conclusion and future work.

II. Methods

We conducted a retrospective observational cohort study using trauma registry data from a Level I trauma center in Kaohsiung, Taiwan, covering the period from January 2023 to December 2024. Inclusion criteria were: Injury Severity Score (ISS) >15, Emergency Department (ED) temperature 32.0-37.9°C, and hospital admission. Trauma patients with a temperature recorded on arrival in the ED between 32.0°C and 37.9°C were included and categorized into two groups: normal temperature (35.0–37.9°C) and hypothermia (32.0–34.9°C). Descriptive statistics were used to summarize patient characteristics. Categorical variables were compared between groups using chi-square tests or Fisher's exact tests, as appropriate. Continuous variables were analyzed using independent t-tests. A p-value <0.05 was considered statistically significant. All analyses were performed using IBM Statistical Package for the Social Sciences (IBM SPSS) Statistics version 25. No advanced predictive algorithms were applied; analyses were limited to standard statistical tests for categorical and continuous variables.

III. Results

Of the 1,324 patients included (ED temperature 32.0–37.9°C), 31 (2.34%) were classified as hypothermic (32.0–34.9°C), while the remainder were normothermic (35.0–37.9°C). The mean age of the hypothermia group was 56.32 ± 22.16 years old, and 51.1% were male. Most of them were transferred patients (90.3%), and 93.5% were triaged as level 1. Trauma team activation occurred in 67.7%, and 35.5% had a Systolic

 ISBN: 978-1-68558-312-5

Blood Pressure (SBP) <90 mmHg. Notably, 93.5% were coma (Glasgow Coma Scale, GCS<9). Over half (54.8%) underwent surgery. The most common injury mechanisms were traffic accident (74.2%) and falls (19.4%). Hypothermia peaked in April (19.4%) and September (16.1%), with autumn being the most frequent season (35.5%).

TABLE I. BASELINE CHARACTERISTICS IN SEVERELY INJURED PATIENTS ON HOSPITAL ARRIVAL

Variable	Mortality N(%)	Survival N(%)	p-value
Sex			0.605
Male	10(55.6)	8(44.48)	
Female	6(46.2)	7(53.8)	
Arrival Method			0.583
EMS	2(66.7)	1(33.3)	
Transfer	14(50.0)	14(50.0)	
Triage			0.131
Level1	16(55.2)	13(44.8)	
Level2	0(0)	2(100)	
SBP(mmHg)			0.31
<90	7(63.6)	4(36.4)	
>=90	9(45.0)	11(55.0)	
GCS			0.131
<9	16(55.2)	13(44.8)	
>=9	0(0)	2(100)	
Season			0.364
Spring	4(44.4)	5(55.6)	
Summer	1(33.3)	2(66.7)	
Autumn	8(72.7)	3(27.3)	
Winter	3(37.5)	5(62.5)	

TABLE II. SEVERITY OF EACH BODY REGION, ICU LOS AND THE OUTCOME OF HYPOTHERMIA

Variable	Mortality mean(SD)	Survival mean(SD)	p-value
Max AIS			
Head/Neck	3.38(1.63)	3.0(2.20)	0.592
Face	0.63(0.89)	0.60(0.91)	0.939
Thorax	2.31(1.74)	2.13(1.81)	0.781
Abdomen	2.25(1.81)	1.07(1.67)	0.068
Extremity	1.69(1.45)	1.067(1.22)	0.209
External	0.94(0.44)	0.93(0.26)	0.975
ISS	33.38(8.41)	28.47(9.49)	0.138
ICU LOS	2.69(2.46)	15.80(8.45)	<0.005*

The mean ISS was 31.0 ± 9.15. The overall mortality rate in the hypothermia group was 51.6%. GCS, Intensive Care Unit Length Of Stay (ICU LOS), and mortality were significantly associated with hypothermia ($p < 0.005$), while no significant associations were found for sex, arrival method, injury mechanism, season, or blood pressure.

IV. DISCUSSION

In southern Taiwan's tropical climate, hypothermia still occurred in major trauma patients and was strongly linked to mortality (51.6%). Although incidence (2.34%) was lower than in temperate (12.6%) and subtropical (5.7%) regions, outcomes were consistent with global benchmarks such as American College of Surgeons-Trauma Quality Improvement Program (ACS-TQIP). This highlights that climate affects incidence but not prognostic significance. Hypothermia remains a universal threat in trauma care, requiring early detection and management. International guidelines, including Tactical Combat Casualty Care (TCCC), recommend passive and active warming, warmed fluids, and minimizing heat loss. Future work should assess the implementation of preventive warming, especially during interhospital transfers, to improve survival outcomes.

V. LIMITATIONS

This study was conducted in a single trauma center, limiting generalizability. Temperature was measured via tympanic thermometry, not core temperature. Most patients were interhospital transfers, and prehospital warming measures could not be verified. The study also lacked data on transfusion status and laboratory values.

VI. CONCLUSION AND FUTURE WORK

Hypothermia can occur even in tropical regions and is significantly associated with mortality in severely injured trauma patients. Although this study focused on trauma patients in a tropical climate, future research should investigate mechanisms to prevent poor outcomes associated with hypothermia. Strategies may include both passive warming (thermal blankets, heated environment) and active warming (warmed intravenous fluids, warming devices), as well as minimizing heat loss during interhospital transfers. Moreover, the methodology used in this study can be replicated in other settings, including temperate and subtropical climates, and may also be extended to other critical conditions such as burns, where hypothermia has prognostic implications.

REFERENCES

[1] L. M. Aitken, J. K. Hendrikz, J. M. Dulhunty, and M. J. Rudd, "Hypothermia and associated outcomes in seriously injured trauma patients in a predominantly sub-tropical climate," Resuscitation, vol. 80, no. 2, pp. 217–223, 2009. https://doi.org/10.1016/j.resuscitation.2008.10.021

[2] G. K. Luna, R. V. Maier, E. G. Pavlin, D. Anardi, M. K. Copass, and M. R. Oreskovich, " Incidence and effect of hypothermia in seriously injured patients," J Trauma, vol. 27, no. 9, pp. 1014-8, 1987. doi: 10.1097/00005373-198709000-00010. PMID: 3656463.

[3] A. M. Jose et al., "Hypothermia on admission predicts poor outcomes in adult trauma patients,". Injury, vol. 56, no. 5, 2024. doi: 10.1016/j.injury.2024.112076. Epub 2024 Dec 3. PMID: 39658434.

[4] M. Azarkane, T. W. H. Rijnhout, I. A. L. van Merwijk, T. N. Tromp, and E. C. T. H. Tan, "Impact of accidental hypothermia in trauma patients: A retrospective cohort study," Injury, vol. 55, no. 1, 2024. doi: 10.1016/j.injury.2023.110973. Epub 2023 Aug 4. PMID: 37563046.

[5] B. L. Bennett et al., "Management of Hypothermia in Tactical Combat Casualty Care: TCCC Guideline Proposed Change 20-01 (June 2020) ," J Spec Oper Med, vol. 20, no. 3, pp.21-35, 2020. doi: 10.55460/QQ9R-RR8A. PMID: 32969001.

[6] Central Meteorological Administration, Ministry of Transportation and Communications (2025/6/30). *113th Annual Climate Data Report - Ground Data.* https://www.cwa.gov.tw/V8/C/D/publication.html?key=5

From Hospitals to Researchers: A Data-Trustee Infrastructure to Search and Use FHIR-Data for Retrospective Medical Research

Carolin Poschen
Department of Computer Science
Trier University of Applied Sciences
Trier, Germany
email: c.poschen@hochschule-trier.de

Joscha Grüger
Experience-based Learning Systems
German Research Center for Artificial Intelligence
Trier, Germany
email: joscha.grueger@dfki.de

Britta Berens
Department of Computer Science
Trier University of Applied Sciences
Trier, Germany
email: b.berens@hochschule-trier.de

Helene Christ
BI & Analytics
Dedalus HealthCare GmbH
Trier, Germany
email: helene.christ@dedalus.com

Lukas Meyer
BI & Analytics
Dedalus HealthCare GmbH
Trier, Germany
email: lukas.meyer@dedalus.com

Konstantin Knorr
Department of Computer Science
Trier University of Applied Sciences
Trier, Germany
email: k.knorr@hochschule-trier.de

***Abstract*—The secondary use of clinical data is crucial for advancing medical research, yet it remains challenged by data fragmentation, privacy concerns, and limited availability. This paper presents a data-trustee infrastructure designed to enable secure, privacy-preserving access to retrospective medical data stored in Hospital Information Systems (HIS). The infrastructure leverages Fast Healthcare Interoperability Resources (FHIR) standards to ensure interoperability and employs a modular pipeline. The pipeline extracts, preprocesses, encrypts, and annotates the data in the hospital and then stores data in a trustee repository. A central component—the Study Specification Board—facilitates ethical and formalized study planning, while a privacy-preserving, two-phase search mechanism allows researchers to retrieve relevant data without exposing sensitive information. A demonstrator system has been implemented and successfully integrated with an HIS, confirming the feasibility and practical applicability of the approach. This work represents a significant step toward operationalizing secure clinical data sharing aligned with EU-GDPR and the goals of the European Health Data Space.**

Keywords-medical research; data-trustee infrastructure; data access; privacy; security.

I. Introduction

In recent years, Artificial Intelligence (AI) has brought significant and disruptive changes across various sectors, including healthcare [1]. However, in the medical field, the integration of AI has so far achieved only limited success [2]. This can be partially attributed to the substantial data requirements for training effective AI models [1], which pose a challenge in healthcare due to the sensitive nature of medical data and associated privacy concerns. Additionally, technical barriers persist: medical data are typically generated and stored across numerous hospitals and private practices, resulting in heterogeneous and fragmented data silos with inconsistent formats and limited interoperability [3]. Serving the objectives of the EU Data Strategy, the European Health Data Space strives to unify these fragmented data silos, relying on technical and semantic interoperability [4]. These data spaces would also provide access to large and robust datasets, which are crucial to train AI models [5].

The secondary use of clinical data is increasingly valued as a vital tool for enhancing healthcare and advancing medical research. Using clinical data for medical research offers several key advantages. Since the data are already collected during routine patient care, they are readily available, cost-effective, and eliminate the need for additional patient involvement or physical intervention. This real-world data enables large, diverse sample sizes, making it especially valuable for studying rare diseases [6]. However, secondary analysis of raw health records poses significant challenges, as the data were initially collected for clinical care rather than research. Researchers must navigate fragmented databases, inconsistent representations of clinical concepts, and changes in coding practices over time, all of which complicate data access and preparation [7]. To overcome these challenges and to ensure secure, trustworthy, and legally compliant access to health data, the concept of data trustees has been proposed [8]. Additionally, data trustees can serve as data spaces as proposed by the European Union and outlined above.

This paper introduces a secure data pipeline and a Data-Trustee Infrastructure (DTI) designed to facilitate privacy-compliant secondary use of medical data. Our approach enables the controlled transfer of data from hospitals to researchers through a data trustee, an intermediary that manages and forwards data without having direct access to its contents. Within our pipeline, medical data are collected and preprocessed securely within the hospital's internal infrastructure, then encrypted and stored in a central repository. Descriptive metadata for each data entry are created to keep a general description while storing the original data encrypted. Researchers may access these data only for specific studies that have received approval from an ethics committee. An automated process translates the approved study's data requirements into search parameters and queries the describing data set. Access to the

original, encrypted data and their corresponding keys is granted only if the number of matching records exceeds a predefined threshold, ensuring both data utility and privacy protection.

The remainder of the paper is structured as follows: Related work in the fields of sharing and accessing medical data for research is discussed in Section II. Section III details the architecture of our DTI and the complete pipeline for data in the system, as well as how researchers interact with it to access data. A discussion of our work follows in Section IV, Section V concludes the paper and outlines potential future work.

II. Related Work

When sharing and using medical data for research, it is important to balance the opportunity provided by data against the individual's right to control their own data [9]. With that in mind, [10] presents a review of research on patient perspectives regarding data sharing, covering their motivations, concerns, privacy considerations, and conditions for sharing. Druedahl et al. concluded that hearing patients' voices is crucial for public acceptance, inclusion, and equity in data sharing.

The German Medical Informatics Initiative (MII) [11] established a decentralized, FHIR-based, federated research data infrastructure based on local Data Integration Centers (DIC) at university centers and partner locations, which extract, pseudonymize, and harmonize clinical data using a modular core dataset defined with international standards. Analyses are performed either centrally—based on a harmonized Broad Consent—or via federated learning, where containerized algorithms are distributed to local sites (data-in-place approach). The German Research Data Portal for Health (FDPG) serves as a central entry point for researchers, offering metadata browsing, feasibility queries, and cohort selection. Though different research projects have already requested data through the MII infrastructure, their data application process still requires substantial manual rework and communication between DIC, FDPG staff, and data requesters as described in [12]. Our proposal minimizes the manual rework and communication overhead by storing data centrally, reducing the number of involved parties for data requests and a simplified and intuitive data requirement description.

A data-trustee architecture for medical sleep research data is presented in [13]. Their architecture enables secure, decentralized data sharing based on dynamic patient consent. A key feature of their system is a standardized, FHIR-based feasibility query that allows researchers to search for relevant data before submitting formal access requests. Combined with containerized analysis environments and tamper-proof logging, the platform addresses legal, ethical, and technical challenges in secondary data use. In addition to addressing these challenges, our approach relies on a separate but centralized data storage architecture, aiming to automate as many steps as possible.

The concept of data trusts or data trustees is discussed in different works. While [14] argues for a variety of data trusts, so that data subjects can choose the most suitable one, [15] aims to answer the question "What are design features that assist practitioners in the secure and sovereign selection process of finding a data trustee in a data space?". When designing data trustee models, [16] identifies four ideal-typical archetypes for data trustees in healthcare, namely data brokerage, processing, aggregation, and custody trustees, which differ along their defined meta-dimensions (1) Task & People, (2) Technology, and (3) Structure.

In [17], the authors propose data trusts as a service using blockchain, which they claim may enable transparent data sharing between multiple stakeholders. To share electronic medical records of the same patients between different hospitals, [18] proposes a blockchain-based information system, MedBlock, as an efficient and privacy-preserving scheme to share data between hospitals. However, as [19] points out in their discussion on leveraging blockchain for healthcare data management systems, the integration of blockchain with healthcare systems generates some challenges, such as interoperability, complexity, or integration with existing systems.

Many works discuss the use of Electronic Health Records (EHR) for medical research [7], [20]–[22]. For example, [20] explored challenges and opportunities of sharing and reusing EHR data for clinical research during the COVID-19 pandemic. They highlight limited syntactic and semantic interoperability, regional privacy regulations, and emerging data protectionism as key barriers. To address privacy regulations and prevent uncontrolled data use, they emphasize the role of a data steward who enforces policies to support institutions in overseeing data sharing both legally and comprehensively. To enable retrospective analysis using EHR data, [7] presents a seven-step data preparation workflow, ranging from obtaining an overview of available data, over extracting relevant data, to implementing a data processing pipeline. Although their work discusses different issues regarding the access and preparation of data for secondary use, it is based on the experience of a single hospital, and the workflow would need potential adjustment for different hospitals. Likewise, [21] proposes an automated framework to transform clinical data into Findable, Accessible, Interoperable, Reusable (FAIR) research data. The implementation targets a maximum-care university hospital, yet, as in prior cases, remains institution-specific and may require adaptation for broader applicability. Similar to our approach, [22] proposes a pipeline to convert EHR data into FHIR standard to support AI research. Their workflow comprises five steps: querying hospital databases, mapping data to FHIR, validating the output, transferring it to a database, and exporting it in an AI-friendly format. However, the authors do not address anonymization or pseudonymization, and instead store all data in plain text within a single database.

This work is an extension to [23]. The previous work focused on the architecture design of the main DTI-components. In contrast, this work focuses on the data flow through the pipeline and the researchers' interaction with the system, including the components in the hospital's and researcher's network.

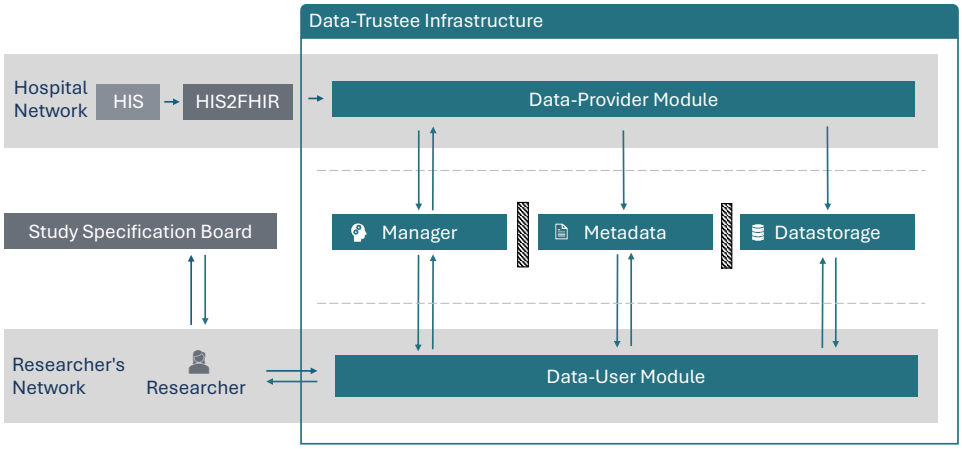


Figure 1. Architecture of the Data-Trustee Infrastructure, as well as the components in the hospital network and the SSB. The arrows indicate data flow between hospital systems, the data-trustee modules, the SSB, and researchers.

III. The Data Pipeline through Our Data-Trustee Infrastructure

The data pipeline through our data-trustee infrastructure, whose architecture was first proposed in [23] and shown in Figure 1, starts in a **Hospital**, where medical data are created during patient care. These data are transformed into the Health Level 7 Fast Healthcare Interoperability Resources (HL7 FHIR) format by the **HIS2FHIR** module that sends data to the **Data-Provider Module**. This module, still deployed in the hospital's local network, carries out further preprocessing, i.e., splitting, encrypting, and annotating data entries. Split data are sent to and stored in the three independent **core modules** of the DTI. A **Study Specification Board** (SSB) helps researchers to formally define study requirements, especially cohort definitions. The SSB operates independently of the DTI and is deployed externally. Though it supports the usage of the DTI through the provision of *Study Specification Documents* (SSDs), it can be used independently of our DTI. The created SSD is used by the **Data-User Module** to carry out a two-phase, three-party, privacy-preserving search. Upon successful completion of this search, researchers obtain relevant data to conduct their retrospective research.

Any interaction between participating units must be operated on a legal basis, such as contracts and compliance.

A. Data Format and Metadata

At all stages of the DTI, medical data are stored and processed in the HL7 FHIR format, the leading interoperability standard for data exchange in healthcare [24]. The FHIR format consists of different resources, like `Patient` or `Encounter`, which contain the relevant information for the specific resource. The `Encounter` resource stores information on the period, type, class, status, and diagnoses of an encounter, while the `Patient` resource contains demographic and identifying information of a specific patient.

During a preprocessing step in our proposed pipeline, medical data are split into data entries, each containing one resource along with its corresponding information. Each medical entry is annotated with descriptive information, referred to as *Metadata*, which provide generalized and categorized information about the original entry. Such data classification can be the storage of an interpretation of a value, e.g., high blood pressure taking factors such as age and sex into account, instead of the actual numerical value. Generalization may involve taxonomy generalizations such as truncating ICD-10 codes, e.g., storing E10 instead of E10.1. These steps allow us to encrypt the original medical data while maintaining a general description of data in an unencrypted format.

B. Study Specification Board

The interaction of a researcher with the DTI is driven by the need to efficiently access retrospective patient data. To get access to those data, researchers first input information about their planned study including their formalized study parameters into the *Study Specification Board (SSB)*. This also includes a positive vote of an ethics committee on their planned study, in accordance to European Union - General Data Protection Regulation (EU-GDPR) Recital 33, which states that sensitive data must be used in compliance with recognized ethical standards. In addition, some national laws require ethics approval before data access is granted.

The formalization of study parameters, previously suggested in [25] and [26], aims to simplify the definition of cohorts, which can then be used for the search of data in a given database, and the subsequent publication of the study. The SSB is designed to provide a user-friendly interface that does not require knowledge of the FHIR format, but rather provides an intuitive approach of data formalization as proposed in [25]. The formalization is based on metadata.

Once the study proposal and its ethics vote have been evaluated in the SSB, the study is published on the SSB. Additionally, a *Study Specification Document (SSD)* is generated that transforms formalized study parameters into a FHIR-compliant format. The SSD forms the basis for delegating the search of relevant data (cf. Section III-F).

C. Hospital

Digital data recorded in a hospital are stored in the HIS as part of the treatment. For secondary use of data, additional consent is required, therefore Broad Consent [27] is used for consent acquisition. This allows for the storage and secondary use of patient data, without tying the consent to a specific study. Patients are approached during discharge. This allows for the consent process to be integrated into the existing workflow and places the patient in a less stressful situation to consider data donation. The consent document is machine-readable, enabling automatic conversion into the FHIR format.

Once consent is given, a patient's data can be extracted for secondary use by the HIS2FHIR module. To standardize the data, semantic mappings to established medical terminologies are introduced. For example, laboratory values are mapped to LOINC. Data themselves are collected on case level and transformed into the FHIR format. It ensures syntactical correctness and enables the merging of data from different hospitals, although this procedure does not consider the semantic correctness of the data. Regarding the secondary use of data, completed cases ensure data consistency, since they are not likely to change retrospectively. Therefore, only discharged cases are considered. Every case of a patient within a hospital, for which consent has been provided, is extracted. In every subsequent export, only the most recent completed case is extracted, minimizing the extracted data per export. Extraction is scheduled, ensuring it occurs consistently at the same daily time. The extracted data are further processed by the Data-Provider Module.

D. Data-Provider Module

The Data-Provider Module (DPM) is the entry point to the DTI for a hospital. It is uniquely configured for each hospital, with its digital identity directly embedded, and deployed in its local network. It receives medical data as FHIR bundles from the HIS2FHIR module. These FHIR bundles are split into demographic and medical data, and the patient's consent document. Demographic data (DDAT) consist of information stored in the resource type `Patient`, while medical data (MDAT) consist of resource types like `Encounter` or `Observation`.

Data stored in the `Patient` resource are pseudonymized by removing all direct identifiers; information such as the patient's gender, birth year, or their truncated postal code are kept. Each patient is assigned a pseudonym to enable privacy-preserving record linkage if new data of the same patient become available [28]. Additionally, the patient resource is enriched with IDs of each corresponding MDAT entry. In the consent document, the identity of the patient is replaced by their pseudonym.

Each MDAT entry is encrypted (eMDAT) using a newly created, unique Data Encryption Key (DEK). The DEK is a symmetric key of a state-of-the-art cryptographic scheme. To ensure searchability of encrypted medical data while aligning with the formalized study parameters defined in the SSB, a metadata record is created for each MDAT entry as described in Section III-A. Thus, each eMDAT entry and its corresponding metadata entry are assigned the same newly generated unique identifier to ensure consistent linkage. Finally, the DPM sends all documents to their corresponding DTI-core modules.

This preprocessing step is performed within the hospital's local network, yet inside the DTI. This enables a secure split, encryption, and annotation of data prior to their storage in dedicated and physically separated modules. The split is essential for the privacy-preserving design of the DTI and data are only merged again by the researcher.

E. DTI-Core Modules

The DTI at its core consists of three modules, the **Manager** module, the **Metadata** module, and the **Datastorage** module. The DTI operates under a strict separation-of-concerns model: no single module has access to both patient identity and medical content. This architectural principle ensures that sensitive associations can only be reconstructed by the authorized researcher within the Data-User Module. They primarily function as independent storage modules with minimal business logic. However, whenever a request is processed, all steps are authorized, and all returned data are signed with a digital signature of the respective module. This ensures authenticity and integrity of results, particularly when they are forwarded to other modules.

The Manager module consists of three different services, each responsible for a different purpose. The *identity service* stores patient pseudonyms together with their DDAT and MDAT IDs, DEKs are stored in the *key service*, and the consent in the *consent service*. The Manager module enables the search on DDAT based on the given consent and returns all MDAT IDs of patients that fit the search criteria, as further described in Section III-F. After a successful search, the Manager module also provides the corresponding DEKs for all found MDAT IDs and issues a signed receipt of all downloadable eMDAT.

The Metadata module stores metadata provided by multiple DPM and enables querying, allowing searches for relevant metadata. Similarly, the Datastorage module stores eMDAT and returns them for given IDs obtained by the search, provided that the present signed receipt is verified as issued by the Manager module.

F. Data-User Module

The Data-User Module (DUM) is the entry point for the researcher to the DTI. Following a successful evaluation of a study proposal within the SSB, a dedicated instance is uniquely generated for the approved research purpose and made available to the researcher. The SSB exports all formal cohort definitions as a Study Specification Document (SSD), a FHIR-compliant format that is directly embedded in the DUM. The SSD must not be modified; otherwise, the entire module is invalidated. This is enforced by integrity-preserving measures, using digital signatures. Once integrity is ensured, the SSD is used to carry out a privacy-preserving, two-phase, three-party search, first proposed in [29] and formalized in [23]. In the first phase, each SSD cohort definition is split into multiple search queries. Queries related to patient demographics are sent to the Manager

module, which returns the MDAT IDs of patients that match the query criteria. Simultaneously, queries concerning medical data are sent to the Metadata module, which responds with IDs of matching MDAT entries. All returned IDs are grouped by patients, and the system determines which patients match, i.e., meet the entire set of specified criteria. In the second phase of the search algorithm, the eMDAT and their DEKs, along with the DDAT of matching patients are requested from the Datastorage and Manager modules respectively. Within the DUM, eMDAT are decrypted and can be used by the researcher.

The search procedure is designed to be privacy-preserving by enforcing a strict separation of data domains. Only the authorized researcher is able to reconstruct the linkage between demographic and medical data. The Manager module operates exclusively on pseudonymized demographic data and associated identifiers without access to any clinical content. In contrast, the Metadata module processes generalized medical metadata without knowledge of patient identities. At no point can either module independently infer complete patient-level information, thereby preventing re-identification risks, while only providing data specifically for a study upholds data minimization.

IV. Discussion

Our proposed DTI offers a practical and privacy-preserving approach that facilitates the secondary use of clinical data for retrospective medical research. A demonstrator implementing the system design has been developed and evaluated, integrated with an HIS, confirming its feasibility and suitability for practical use in clinical research environments.

A. Strengths and Contributions

A key contribution of this work is the development of a nearly fully-automated pipeline that enables the secure transfer of patient data from hospitals to researchers. By leveraging a modular, standardized architecture based on HL7 FHIR, we enhance interoperability and reduce the technical integration burden across institutions. The process—from in-hospital data preprocessing, encryption, and metadata annotation, to study-specific data retrieval by researchers—is handled in a streamlined, privacy-conscious manner.

Our infrastructure supports researchers in accessing initially distributed datasets via a unified system. The SSB and DUM simplify study setup and automate the formalization and translation of cohort definitions into FHIR-compatible search parameters. This ensures legal and ethical compliance (e.g., with EU-GDPR and ethics committee approval) and reduces researcher workload and administrative overhead.

Furthermore, the two-phase, three-party, privacy-preserving search mechanism ensures that the DTI-core modules cannot infer sensitive links between patient identities and medical content. Only the researcher, within their working environment, can decrypt and reconstruct the data necessary for their approved study.

B. Limitations and Challenges

Despite these strengths, several limitations remain that could affect scalability and adoption. First, participation from hospitals requires technical integration efforts, including the deployment of specific components such as customized HIS2FHIR and DPM. Secondly, each research project requires its own DUM instance. Although this leads to a certain amount of additional work, it is limited in time, as the DUM can be taken out of operation again once the data has been delivered.

Another practical limitation is the manual verification of actors and identities at onboarding. While this step is common across most trusted data-sharing ecosystems, it remains a bottleneck and may benefit from future integration with national digital identity systems.

Moreover, the system currently depends on metadata for search operations. While this approach supports general cohort definitions and preserves privacy, it limits the granularity and specificity of data queries. Highly specialized or narrow study parameters may not be captured by available metadata alone.

Finally, data accessed through the DTI are not fully anonymized. Although encryption, access control, and legal contracts serve as safeguards against re-identification, the lack of guaranteed anonymization represents a residual privacy risk that must be addressed through governance and compliance measures.

V. Conclusion and Future Work

This work provides a concrete and extensible blueprint for operationalizing the principles of the European Health Data Space. It tackles key challenges such as patients' consent and heterogeneous data formats. A two-phase, three-party, privacy-preserving search algorithm guarantees that the patients' data can only be combined by the researcher. This ensures that the other parties cannot access the data, and the researchers are only able to use data they have permission to. While the data are stored centrally, they are split into distinct components. This design allows searches to be performed solely on the metadata, completely isolating the encrypted raw data from the querying process. Furthermore, we automated the entire processes of getting data from hospitals to requesting data for research, thereby eliminating a bottleneck in retrospective medical research.

Future enhancements could include:

- Integration of outpatient care data and general practitioners. This can be achieved by deploying a module similar to the HIS2FHIR component in the practitioner's system.
- Support for the re-import and analysis of research outcomes to promote learning healthcare systems.
- Semantic enrichment of data and improved quality checks to ensure plausibility and consistency.
- Mechanisms for patient-driven consent management and dynamic revocation.
- More expressive query languages for SSDs, potentially combined with privacy-preserving computation techniques like secure multi-party computation or federated analytics.
- Systematic and quantitative evaluation of the DTI.

Overall, while challenges remain, our infrastructure represents a significant step toward bridging the gap between clinical

data silos and the data needs of modern AI-driven healthcare research.

ACKNOWLEDGEMENT

This work is part of the DaTreFo project, funded by the German Federal Ministry of Research, Technology, and Space (16KIS1644).

REFERENCES

[1] A. B. Rashid and A. K. Kausik, "AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications", *Hybrid Advances*, p. 100 277, 2024. DOI: 10.1016/j.hybadv.2024.100277.

[2] A. Zahlan, R. P. Ranjan, and D. Hayes, "Artificial intelligence innovation in healthcare: Literature review, exploratory analysis, and future research", *Technology in Society*, vol. 74, p. 102 321, 2023. DOI: 10.1016/j.techsoc.2023.102321.

[3] T. K. Eisinger-Mathason *et al.*, "Data linkage multiplies research insights across diverse healthcare sectors", *Communications Medicine*, vol. 5, no. 1, p. 58, 2025. DOI: 10.1038/s43856-025-00769-y.

[4] C. Stellmach, M. R. Muzoora, and S. Thun, "Digitalization of Health Data: Interoperability of the Proposed European Health Data Space ", in *Digital Professionalism in Health and Care: Developing the Workforce, Building the Future*, IOS Press, 2022, pp. 132–136. DOI: 10.3233/SHTI220922.

[5] I. Ulnicane, "Artificial Intelligence in the European Union: Policy, ethics and regulation", in *The Routledge Handbook of European Integrations*, Taylor & Francis, 2022. DOI: 10.4324/9780429262081-19.

[6] M. Jungkunz, A. Köngeter, K. Mehlis, E. C. Winkler, and C. Schickhardt, "Secondary Use of Clinical Data in Data-Gathering, Non-Interventional Research or Learning Activities: Definition, Types, and a Framework for Risk Assessment", *Journal of Medical Internet Research*, vol. 23, no. 6, e26631, 2021. DOI: 10.2196/26631.

[7] A. Maletzky *et al.*, "Lifting Hospital Electronic Health Record Data Treasures: Challenges and Opportunities", *JMIR Medical Informatics*, vol. 10, no. 10, e38557, 2022. DOI: 10.2196/38557.

[8] S. Kilz and M. Radic, "Health Data Trustees: A Business Model Perspective", in *The International Conference on Innovations in Computing Research*, Springer, 2024, pp. 618–630.

[9] T. Hulsen, "Sharing Is Caring—Data Sharing Initiatives in Healthcare", *International Journal of Environmental Research and Public Health*, vol. 17, no. 9, p. 3046, 2020. DOI: 10.3390/ijerph17093046.

[10] L. C. Druedahl and S. Kälvemark Sporrong, "Patient Perspectives on Data Sharing", in *The Law and Ethics of Data Sharing in Health Sciences*, Springer, 2023, pp. 51–67. DOI: 10.1007/978-981-99-6540-3_4.

[11] S. C. Semler *et al.*, "The Medical Informatics Initiative at a glance-establishing a health research data infrastructure in Germany", *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz*, pp. 616–628, 2024. DOI: 10.1007/s00103-024-03887-5.

[12] H.-U. Prokosch *et al.*, "Towards a National Portal for Medical Research Data (FDPG): Vision, Status, and Lessons Learned", in *Caring is Sharing — Exploiting the Value in Data for Health and Innovation*, IOS Press, 2023, pp. 307–311. DOI: 10.3233/SHTI230124.

[13] R. Burmeister, C. Erler, F. Gauger, R. J. Dressle, and B. Feige, "Advancing Sleep Research Through Dynamic Consent and Trustee-Based Medical Data Processing", ICDS, 2024, ISBN: 978-1-68558-169-5.

[14] S. Delacroix and N. D. Lawrence, "Bottom-up data Trusts: disturbing the 'one size fits all' approach to data governance", *International Data Privacy Law*, vol. 9, no. 4, pp. 236–252, 2019. DOI: 10.1093/idpl/ipz014.

[15] M. Steinert, D. Tebernum, and M. Hupperz, "Design Features for Data Trustee Selection in Data Spaces", in *International Conference on Data Science, Technology and Applications 2024*, 2024, pp. 559–570. DOI: 10.5220/0012851400003756.

[16] F. Lauf *et al.*, "Exploring Design Characteristics of Data Trustees in Healthcare - Taxonomy and Archetypes", *ECIS 2023 Research Papers*, p. 323, 2023.

[17] R. K. Lomotey, S. Kumi, and R. Deters, "Data Trusts as a Service: Providing a platform for multi-party data sharing", *International Journal of Information Management Data Insights*, vol. 2, no. 1, p. 100 075, 2022. DOI: 10.1016/j.jjimei.2022.100075.

[18] K. Fan, S. Wang, Y. Ren, H. Li, and Y. Yang, "Medblock: Efficient and Secure Medical Data Sharing Via Blockchain", *Journal of Medical Systems*, vol. 42, pp. 1–11, 2018. DOI: 10.1007/s10916-018-0993-7.

[19] I. Yaqoob, K. Salah, R. Jayaraman, and Y. Al-Hammadi, "Blockchain for healthcare data management: Opportunities, challenges, and future recommendations", *Neural Computing and Applications*, pp. 1–16, 2022. DOI: 10.1007/s00521-020-05519-w.

[20] A. Dagliati, A. Malovini, V. Tibollo, and R. Bellazzi, "Health informatics and EHR to support clinical research in the COVID-19 pandemic: an overview", *Briefings in Bioinformatics*, vol. 22, no. 2, pp. 812–822, 2021. DOI: 10.1093/bib/bbaa418.

[21] M. Parciak *et al.*, "FAIRness through automation: development of an automated medical data integration infrastructure for FAIR health data in a maximum care university hospital", *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, p. 94, 2023. DOI: 10.1186/s12911-023-02195-3.

[22] E. Williams *et al.*, "A Standardized Clinical Data Harmonization Pipeline for Scalable AI Application Deployment (FHIR-DHP): Validation and Usability Study", *JMIR Medical Informatics*, vol. 11, e43847, 2023. DOI: 10.2196/43847.

[23] C. Poschen, B. Herres, and K. Knorr, "A Threat-Driven Design of a Data-Trustee Infrastructure for Medical Data", in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2024, pp. 6753–6760. DOI: 10.1109/BIBM62325.2024.10822788.

[24] S. N. Duda *et al.*, "HL7 FHIR-based tools and initiatives to support clinical research: a scoping review", *Journal of the American Medical Informatics Association*, vol. 29, no. 9, pp. 1642–1653, 2022. DOI: 10.1093/jamia/ocac105.

[25] C. Poschen, B. Berens, and K. Knorr, "Towards Formalized Study Parameters for Medical Research", in press, to be published at MCCSIS e-health 2025, 2025.

[26] B. Berens, J. Grüger, C. Poschen, and K. Knorr, "A FHIR Specification to Formalize Cohort Definitions", in press, to be published at EFMI Special Topic Conference 2025 Good Evaluation - Better Digital Health, 2025.

[27] D. Hallinan, "Broad consent under the GDPR: An optimistic perspective on a bright future", *Life Sciences, Society and Policy*, vol. 16, no. 1, p. 1, 2020. DOI: 10.1186/s40504-019-0096-3.

[28] A. Gkoulalas-Divanis, D. Vatsalan, D. Karapiperis, and M. Kantarcioglu, "Modern Privacy-Preserving Record Linkage Techniques: An Overview", *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4966–4987, 2021. DOI: 10.1109/TIFS.2021.3114026.

[29] B. Herres, C. Poschen, and K. Knorr, "Privacy-Preserving Search on Medical Data", in *Digital Health and Informatics Innovations for Sustainable Health Care Systems*, IOS Press, 2024, pp. 252–256. DOI: 10.3233/SHTI240392.

School Health Dialogue: A Prompt-Expansion and Response-Visualization Framework

Hayato Tomisu†
R-GIRO / Graduate School of Data Science
Ritsumeikan Univ. / Shiga Univ.
Shiga, Japan
e-mail: tomisu@fc.ritsumei.ac.jp

Kazue Yamamura†
Graduate School of Human Science / School Nurse
Ritsumeikan Univ. / Ritsumeikan Moriyama J&SHS
Shiga, Japan
e-mail: kazue926@mrc.ritsumei.ac.jp

Junya Ueda
ImpactLab.
Shiga, Japan
e-mail: jueda@impactlab.jp

Tsukasa Yamanaka
Faculty of Life Sciences
Ritsumeikan Univ.
Shiga, Japan
e-mail: yaman@fc.ritsumei.ac.jp

***Abstract*—Adolescents often express mental or physical discomfort in vague terms, thereby placing a high cognitive burden on school nurses, who must interpret incomplete information. This study proposes a two-layer framework to improve school health communication by transforming ambiguous student utterances into structured, explainable dialogue flows. The first layer, auto-prompt expansion, enriches student input into slot-based representations. The second, the Prompt-Graph Domain-Specific Language, maps these representations onto a transparent decision graph for nurse supervision. The system integrates large language model orchestration, animated avatars, and real-time graph rendering. In an evaluation of 50 student complaints, prompt expansion achieved an F1 score of 0.82, whereas slot extraction scored 0.43 owing to lexical variability. AI-based rubric evaluations yielded high tone scores, indicating consistent empathy in responses; however, lower ratings for accuracy and completeness revealed deficiencies in medical specificity and follow-up guidance. Future studies will address clinical tuning, symptom normalization, and long-term field validation.**

Keywords-large language models; decision graph visualization; conversation management; adolescent wellness; school infirmary.

I. Introduction

Mental health complaints among Japanese junior and senior high school students have grown significantly complex since the onset of the COVID-19 pandemic. However, adolescents often describe their symptoms vaguely, leaving school nurses to make rapid severity decisions based on fragmented information. At Ritsumeikan Moriyama Junior and Senior High School, for example, the total number of infirmary visits increased by 52.6% between 2019 and 2021, with 38% of those visits related to non-injury and non-illness issues [1]. This growing cognitive load and diagnostic pressure highlight the urgent need for tools that transform weak, ambiguous student utterances into actionable clinical cues, while preserving the nurse's supervisory role.

Large Language Model (LLM)-based diagnostic support systems have shown promise in hospital settings but typically assume well-structured input for adult users. These systems are not designed to process the multi-symptom, low-verbal expressions typical of schoolchildren, nor do they provide the transparency required for nurse oversight. Consequently, a significant gap persists at the intersection of school health, adolescent mental health, and explainable Artificial Intelligence (AI). To address this gap, we propose a two-part framework: auto-prompt expansion, which enriches vague student input into structured representations; and Prompt-Graph Domain-Specific Language (DSL), which maps these representations onto explainable diagnostic flows. The system reduces nurses' cognitive burden and improves the consistency of initial assessments.

The remainder of this paper is organized as follows: Section II reviews related work, Section III describes the proposed method, and Section IV details its technical implementation. Section V presents a performance evaluation of the proposed method. Section VI discusses the evaluation results, limitations, and future directions. Finally, Section VII concludes the paper.

II. Related Work

A. Automatic Prompt Expansion for LLM

LLMs showed impressive few-shot abilities when supplied with carefully crafted prompts [2]. To minimize manual effort, a line of research has emerged on automatic prompt engineering. AutoPrompt searches the discrete token space via gradients to elicit factual or sentiment knowledge from masked language models [3]. In contrast, soft-prompt tuning learns continuous prefix embeddings that can be mixed and weighted for downstream tasks [4]. Reinforcement Learning Prompt (RLPrompt) frames prompt tokens as an action space and optimizes them with reinforcement learning rewards [5], while Gradient-free Instructional Prompt Search (GrIPS) performs gradient-free, edit-based instruction search to improve natural-language prompts iteratively [6]. More

recent methods exploit the models themselves: Auto-Chain-of-Thought (CoT) samples its reasoning chains to create enriched demonstrations automatically [7], and Zhou et al. showed that LLM can rival humans at generating and ranking high-quality prompts [8]. Building on these insights, our Auto-Prompt Expansion optimizes prompts for adolescent health utterances.

B. Explainable and Visual Reasoning

Prompting strategies have also been used to expose model reasoning. Chain-of-Thought (CoT) prompting makes LLMs emit intermediate steps, creating human-readable rationales [9]. Accuracy improves when multiple reasoning chains are sampled and the most frequent answer is selected [10]. Least-to-Most prompting decomposes complex problems into ordered sub-tasks, yielding an explicit plan-and-solve trace [11]. The ReAct framework interleaves rationales with executable actions, such as web searches, so each step is transparent [12]. Tree-of-Thoughts extends this idea to a branching search that can be audited after inference [13]. Extending plain text, ReasonGraph visualizes reasoning paths as interactive flow diagrams [14], and GraphReason converts multiple CoT traces into a unified graph to detect contradictions [15]. Self-Refine lets the model critique and iteratively refine its answers, exposing an evolution of thought that humans can inspect [16].

Despite these proposals, existing work has been evaluated mainly on synthetic Question Answering or programming tasks. None targets the ambiguity, safeguarding needs, and workflow constraints of school health communication. The proposed Prompt-Graph DSL addresses this gap by converting enriched student statements into a deterministic dialogue graph rendered for nurse monitoring, unifying automatic prompt expansion with visual explainability.

III. Proposed Method

We aimed to develop an automated AI chatbot that reduces the burden on school nurses by capturing various student inputs and enabling nurses to efficiently evaluate and monitor the chatbots' behavior.

The design is guided by three core principles:

1) *Minimal cognitive load*: Students should be able to convey their condition using brief, natural-language utterances, without requiring specialized prompting.
2) *Human transparency*: Nurses must understand why the model responds in a particular way.
3) *Auditability*: Every reasoning step must be reproducible from stored artifacts; no specialized prompting is required.

To satisfy these principles, we propose a two-layer architecture (Figure 1). Layer 1 automatically expands vague prompts into structured representations. Layer 2 deterministically maps these representations onto a declarative decision graph and renders the selected path for nurse supervision. Collectively, these layers enable:

1) the generation of richer, more informative responses from limited input, and
2) the provision of structured, multi-slot symptom representations for supervising nurses.

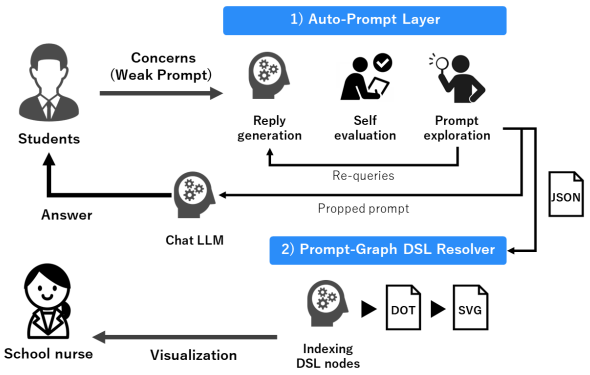


Figure 1. Overview of the proposed method.

A. Problem Statement

Given a short, often vague student utterance u, we aim to produce (i) a nurse-safe reply r, (ii) a slot vector $J = (sym, i, d, c)$, and (iii) a path P on a nurse-auditable Prompt-Graph. The following mathematical expressions do not represent optimization procedures in the implementation; they function as compact design abstractions for prompt engineering. The conceptual search dynamics operate on an explicit state graph, which we model as a beam search over states by drawing inspiration from Grice's cooperative principle in conversation [17]. When information is missing, the system applies the Quantity maxim, which prompts clarification. When the tone is overly rigid, the Manner maxim guides adjustments to improve clarity and communicative ease [18]. In practice, this search is implemented using iterated template prompts; no vector arithmetic or gradient-based optimization is involved.

Each prompt engineering intuition is mapped to three formal objects:

State
$s = \langle u, J \rangle$ — current system prompt u and the partially filled slot vector J.

Operator
Γ = {APPEND_ASK, APPEND_EMPATHY, APPEND_SAFE} — add a clarifying question / add an empathy phrase / prepend a safety reminder.

Score
$\text{Score} = w_c\, C + w_e\, E + w_s\, S$ (1) — weighted sum of Coverage, Empathy, and Safety.

The auxiliary quantities are defined as follows:

$coverage(J)$
$C := coverage(J) = \frac{\#\text{ filled slots}}{4} \in [0,1]$. Full coverage, therefore, yields $C = 1$.

$safe(r)$
$S := safe(r) \in \{0,1\}$, returns 1 if the reply r satisfies all safety filters else 0.

E Cosine similarity between the reply embedding and a fixed "empathy prototype" vector.

w_c, w_e, w_s
Non-negative weights are chosen empirically on a validation set.

τ The score threshold that a candidate must reach to be accepted; tuned once per deployment.

For "My head feels heavy and a little queasy.", the first pass yields sym = headache, i = mild, c = nausea, $d = \varnothing$ ($coverage = 0.75$). Operator ASK requests duration, completes J, and the resolver maps J deterministically to a node P (e.g., mild headache+nausea, duration $<$ 1h).

Our two-layer pipeline realizes a function

$$F : \mathcal{U} \longrightarrow \mathcal{R} \times \mathcal{J} \times \mathcal{P} \tag{2}$$

where $\mathcal{U}$ is the space of student utterances and $\mathcal{R}$ the space of chatbot replies, $\mathcal{J}$ is the space of complete slot vectors, and $\mathcal{P}$ is the set of paths in the Prompt-Graph. This function assigns each utterance $u \in \mathcal{U}$ to a triple (r, J, P), where

- $r \in \mathcal{R}$ is the chatbot's reply,
- $J \in \mathcal{J}$ is the fully populated slot vector,
- $P \in \mathcal{P}$ is the path selected in the Prompt-Graph.

B. Layer 1: Auto-Prompt Expansion

In this layer, the system prompt u and the partial slot vector J are used to update the current state. Here, $sym \in \mathcal{S}$ denotes the symptom, $i \in (\text{mild}, \text{moderate}, \text{severe})$ represents intensity, $d \in \mathbb{R}_{>0}$ indicates duration in minutes, and c denotes a set of co-symptoms. Each state corresponds to an entry in the Knowledge Base, and operator effects are realized through system prompt substitutions.

After each LLM call, the system parses the reply to update the slot vector J and recompute the auxiliary variables C, E, and S. The same acceptance criterion, $\text{Score} \geq \tau$ is applied. Let ψ denote the final sequence of system prompts obtained at termination. Upon success, the layer outputs: (i) the consolidated JavaScript Object Notation (JSON) record $J^\star$ and (ii) the dialogue trace $\langle \psi, r, \text{Score} \rangle^\star$, both of which are passed to Layer 2 for path visualization.

C. Layer 2: Prompt-Graph DSL Resolver

Given $J^\star$, the resolver evaluates predicates in breadth-first order to identify the first matching node. Given that the predicates are mutually exclusive by design, the resolver operates deterministically. Each edge carries either a follow-up question or an action suggestion. The $J^\star \mapsto P$ is logged to support offline replay and error analysis.

The resolver exports the subgraph $G_P = (V_P, E_P)$, induced by all the nodes within two hops (i.e., children and grandchildren) of the current node. The view is updated at every turn, thereby providing nurses with situational awareness while respecting the simplicity constraints of the kiosk environment.

IV. Implementation

Figure 2 illustrates the architecture of the digital school nurse system as deployed in the pilot environment. The system employs a service-oriented design, wherein independent microservices provide conversational intelligence, prompt expansion, and graph visualization. Developed using Python 3.11 and orchestrated by Dify, the system integrates three key open-source components: ChatdollKit for the animated student interface, Dify for LLM orchestration, and a custom visualization module for nurse-side, node-level feedback. This section details the operational runtime cooperation among these services, in accordance with the requirements outlined in Section III.

A. Conversation Pipeline

The conversation pipeline executes as follows:

1) **Student Action**: A student verbally interacts with the digital nurse avatar. The system captures the speech, converts it to text, and transmits it to the Dify chat API.
2) **Prompt Enhancement:** An agent (o3-mini) rewrites the system prompt based on the rules defined in Prompt 1.
3) **Self Evaluation**: The generated response is scored by GPT-4o-mini using the evaluation prompt in Prompt 2. If the score falls below a predefined threshold, steps 2 and 3 are iterated up to three times until the threshold is met.
4) **Response Generation**: Dify sends the enhanced prompt and a system template (Prompt 3) to GPT-4o to produce the final reply.
5) **Slot Extraction**: Using the enhanced prompt, GPT-4o-mini extracts key information into a JSON schema as detailed in Prompt 4. These JSON objects are logged to a JSON Lines (JSONL) file for historical review.
6) **Prompt-Graph Rendering**: The resulting JSON record is forwarded to the Prompt-Graph Resolver, where a Python script converts it to a NetworkX graph via YAML, extracts the relevant subgraph, and renders it as an SVG file using Graphviz.
7) **Output Delivery**: The validated reply is sent back to ChatdollKit to be vocalized by the avatar, while the corresponding SVG visualization is simultaneously embedded into the nurse's monitoring dashboard.

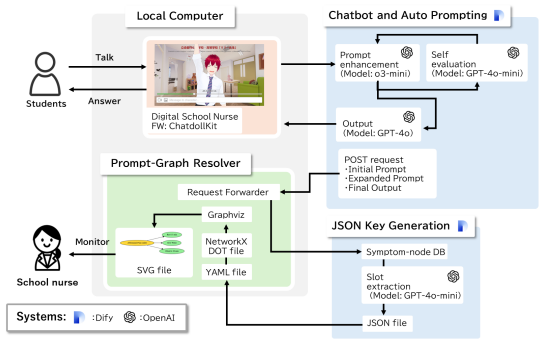


Figure 2. System overview integrating an animated avatar, LLM orchestration, and nurse-facing Prompt-Graph visualization.

```
You are a **Prompt-Expansion Agent**.
Your task is to transform each utterance of health care
     from a Japanese school student into a richer, natural-
     sounding Japanese sentence or short paragraph that
     fully surfaces the student's meaning so later modules
     can answer more accurately.
For every input line:
1. Rewrite it in Japanese, incorporating any information
     that is already present about:
- When it happened, started, or ended
- Where it took place
- What exactly happened or is needed
- Why it is a problem or need
- How it is unfolding or how the student feels
(Include **Who** only if the original utterance mentions
     people explicitly.)
2. Do **not** invent facts or add content that is absent
     from the original utterance.
3. Keep the output plain-no numbering, bullet lists,
     roleplay tags, or JSON.
Return **only** the enhanced Japanese sentence or short
     paragraph.
You must answer in Japanese.
```

Prompt 1: Expansion prompt for utterance enrichment by o3-mini

```
You are a **Prompt-Expansion Agent**.
For each single-line Japanese utterance from a middle-
     school student (``raw input''):

1. Produce an **enhanced Japanese sentence or short
     paragraph** (``enhanced'') that restates the raw input
      and, where the original wording already contains it,
     explicitly includes:
   - When (timing / duration)
   - Where (location)
   - What (event, symptom, request, etc.)
   - Why it matters to the student
   - How the situation is unfolding or how the student
        feels
   (Add **Who** only if it is explicitly mentioned.)
   *Never invent details that are not in the raw input.*
2. **Self-check** the ``enhanced'' text and assign
     quantitative scores (0.00-1.00, two decimals):
   - **coverage** = fraction of the five W-H elements
        present
   - **empathy** = degree of empathy conveyed toward a
        young student
   - **safety** = 1.00 if the text contains **no** medical
         diagnosis, drug names, or prescriptive medical
        advice; otherwise 0.00
   If **coverage < 0.75** **or** **safety < 1.00**,
        rewrite the ``enhanced'' text once, then
        recalculate the scores and use the improved version
        .
3. **Output format (exactly two lines, no extra text)**
   {{ENHANCED}}
   [coverage={{COV}}, empathy={{EMP}}, safety={{SAFE}}]
   Replace the placeholders with the final enhanced text
        and the three numeric scores.
   Do **not** add numbering, bullet points, role-play tags
        , or JSON.
```

Prompt 2: Self-evaluation prompt for prompt quality scoring by GPT-4o-mini

```
You are a friendly school nurse assistant.
Collect four items: symptom, intensity (mild/moderate/
     severe),
duration (minutes or hours), and co-symptoms (comma
     separated).
Ask only one clarifying question at a time.
Avoid medical diagnoses and medication names.
```

Prompt 3: Chatbot prompt for structured response generation by GPT-4o

```
{
"symptom": "headache",
"intensity": "mild",
"duration": "15min",
"coSymptom": "nausea"
}
```

Prompt 4: Slot extraction prompt for filling the JSON schema by GPT-4o-mini

B. *Prompt-Graph Resolver and Visualization*

At startup, it parses the YAML file converted from JSON into a Pydantic object tree and constructs a directed multi-graph using NetworkX. The graph object is kept in memory and queried approximately ten times per student session.

Each traversal event is serialized as $\langle t, J^{\star}, n_{\text{prev}}, n_{\text{next}} \rangle$ and appended to a compressed JSONL file for offline analytics. For visual feedback, the resolver extracts the sub-graph reachable within two hops of the current node, exports it to DOT format, and invokes Graphviz in headless mode to export SVG.

C. *Use Case of Implemented System*

Figure 3 shows the system in use during its in-situ pilot deployment. Positioned within the school infirmary, the kiosk invites students to interact with the animated digital nurse avatar.

V. Performance Evaluation

A. *Procedure*

A corpus of 50 student complaints was prepared: 20 were collected from handbooks and public websites intended for school nurses, and 30 ambiguous examples were mined from social media and public question-and-answer platforms. For each utterance, we created three gold-standard artifacts: (1) an enhanced prompt that rephrased the complaint in a more informative context, (2) an exemplary nurse response, and (3) a four-slot JSON record capturing symptom, intensity,

Figure 3. Students interacting with the deployed system.

duration, and coSymptom. These artifacts were initially corrected using GPT-4o-mini with a simple prompt (shown in Prompt 5) to fix typos, missing characters, and expressions difficult for the model to interpret. The creator then manually verified that the original intent remained unchanged, thereby finalizing the gold-standard dataset.

All model runs followed the configurations shown in Figure 2. Prompt expansion was conducted at a temperature of 0.7, whereas slot extraction used a temperature of 0.1 for higher determinism. Prompt fidelity was evaluated using standard confusion matrix metrics—accuracy, precision, recall, F1 score—where a cosine similarity of ≥ 0.70 (measured using Sentence-Bidirectional Encoder Representations from Transformers (BERT), all-mpnet-base-v2) was considered a true positive. Answer quality was assessed by GPT-4o using a rubric-based prompt (Prompt 6), with scores from 1–5 assigned for accuracy, completeness, and tone. JSON slot extraction was evaluated using exact string matches per slot, with True Positives (TP) for correct values, False Negatives (FN) for missing or incorrect values, and False Positives (FP) for spurious outputs. All utterances used for evaluation were drawn from publicly available materials or anonymized examples; no identifiable student data were collected.

B. Results

1) Prompt Expansion: The enhanced prompts achieved 41 TP and 9 FN, yielding an F1 score of 0.82. No hallucinated slot values were observed below the similarity threshold. This indicated that most errors stemmed from partial omissions rather than from fabrication.

2) Answer Quality: Rubric-based evaluation produced mean scores of 3.63 for accuracy, 3.71 for completeness, and 4.73 for tone on a five-point scale. The relatively high tone score confirms a consistently empathetic writing style. In contrast, the lower accuracy and completeness scores, which were manually verified, revealed gaps in medical specificity and follow-up guidance that require further prompt engineering.

3) JSON Slot Extraction: Using the lower temperature setting, the system produced 39 TP slots, 56 FP slots, and 47 FN slots, which resulted in an F1 score of 0.43. Most of the remaining errors stemmed from lexical variation in location terms and inconsistent mapping of free-text severity phrases to the three-level scale.

```
Please correct typos, missing characters, and expressions
     that may be difficult for machines to interpret. Do
     not change the meaning or intent. Keep the original
     wording as much as possible. The input and output text
      is in Japanese.
```

Prompt 5: Proofreading prompt for output correction by GPT-4o-mini

```
SYSTEM_MSG = (
"You are a strict school nurse evaluator. Given a gold
     reference answer and "
"a candidate answer, rate the candidate on Accuracy,
     Completeness, and Tone "
"on a scale from 1 (poor) to 5 (excellent). "
)
USER_TEMPLATE = (
"GOLD:\n{gold}\n\nCANDIDATE:\n{pred}\n"
)
```

Prompt 6: Rubric-based evaluation prompt for answer scoring by GPT-4o

VI. Discussion

A. Quantitative Findings from Performance Evaluation

The results in Section V indicate that the proposed pipeline effectively captures student intent through prompt expansion by achieving an F1 score of 0.82. This suggests that the model can reliably augment vague student utterances into structured representations when supported by an auto-prompt expansion. Conversely, structured slot extraction remains a significant bottleneck, with an F1 score of 0.43. A detailed error analysis revealed that most FPs and FNs were attributable to lexical variations in symptom and severity expressions. The system attained a precision of 0.41 and a recall of 0.45, suggesting that the current model struggles to consistently identify and normalize the key elements of the complaint. To address this limitation, additional normalization techniques such as controlled vocabularies, synonym clustering, or regular expression filters may be necessary.

Moreover, rubric-based evaluations indicate that while the consistently high tone scores confirm the system's ability to generate empathetic responses, the lower accuracy and completeness scores expose deficiencies in medical specificity and the inclusion of appropriate follow-up guidance.

B. Qualitative Feedback from Pilot Deployment

Qualitative observations during the pilot deployment revealed promising user acceptance and workflow integration. Students reported that the animated avatar was "less intimidating than talking to an adult right away," and many appeared to be more willing to disclose emotional or ambiguous symptoms. This is particularly valuable in the context of adolescent mental health, where emotional or social barriers often hinder verbal expression.

From the school nurses' perspective, the visual decision graph rendered on the Scalable Vector Graphics (SVG)

dashboard enabled passive oversight. This form of transparent feedback preserved the nurse's supervisory role while minimizing cognitive overhead.

Notably, this qualitative evaluation focused on user interaction and interface design rather than on the diagnostic capabilities of the proposed pipeline. The deployed system was a prototype built on GPT APIs, without integration of the auto-prompt expansion or Prompt-Graph resolver described in this study. Therefore, the feedback should be interpreted as a formative assessment of interaction design, not a summative evaluation of system performance.

C. Limitations

This study is a controlled pilot with simulated student inputs, so external validity remains limited. We do not include external baselines in the camera-ready; instead, we release the prompts, scoring rubric, and slot schema to support replication. Slot extraction currently relies on LLM generalization without domain-specific post-editing, hence structured outputs are not yet suitable for unsupervised clinical use. The qualitative evaluation is small and short-term and does not assess long-term behavior or health outcomes. Near-term work targets synonym normalization, ordinal severity mapping, lightweight post-editing rules, a small supervised adapter, and per-slot error analysis.

VII. Conclusion and Future Work

This study introduced a two-layer framework designed to mitigate the cognitive load on school nurses by transforming ambiguous student complaints into structured, explainable dialogue flows. Our evaluation demonstrated that the auto-prompt expansion layer effectively enriches vague inputs, achieving a high F1 score of 0.82. However, the subsequent slot extraction process remains a challenge, with an F1 score of 0.43, highlighting issues with lexical variability in student expressions. While AI-based rubric evaluations confirmed the system's ability to generate empathetic responses, they also revealed deficiencies in medical specificity and follow-up guidance, underscoring the need for further refinement.

Future work will concentrate on enhancing clinical reliability and robustness. Key priorities include the implementation of controlled vocabularies and synonym normalization to improve the precision of slot extraction. Furthermore, a long-term field study is necessary to validate the system's real-world effectiveness, assess its impact on nurse decision-making, and understand how student interactions evolve over time. Through these efforts, we aim to develop a clinically reliable tool that enhances adolescent health support in educational settings.

Acknowledgment

This work was supported by JST RISTEX Japan Grant Number JPMJRS24K3, the Japan Health Foundation, the I-O DATA Foundation, and the Sasakawa Scientific Research Grant from the Japan Science Society. In this work, authors marked with † (Hayato Tomisu and Kazue Yamamura) contributed equally as co-first authors. We would like to thank Editage (www.editage.jp) for English language editing.

References

[1] K. Yamamura, "Current Situation of Visits to the Health Office during the Corona Disaster （コロナ禍における保健室来室の現状）," *Journal of the Japanese Association for the Study of Guidance*, vol. 40, pp. 11–17, 2023.

[2] T. B. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

[3] T. Shin *et al.*, "AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts," in *Proc. EMNLP*, 2020, pp. 4222–4235.

[4] G. Qin and J. Eisner, "Learning How to Ask: Querying Language Models with Mixtures of Soft Prompts," in *Proc. NAACL*, 2021, pp. 5203–5212.

[5] M. Deng *et al.*, "RLPrompt: Optimizing Discrete Text Prompts with Reinforcement Learning," in *Proc. EMNLP*, 2022, pp. 9593–9611.

[6] A. Prasad *et al.*, "GrIPS: Gradient-Free, Edit-Based Instruction Search for Prompting Large Language Models," in *Proc. EACL*, 2023, pp. 3083–3099.

[7] Z. Zhang *et al.*, *Automatic chain of thought prompting in large language models*, arXiv:2210.03493, 2023.

[8] Y. Zhou *et al.*, *Large language models are human-level prompt engineers*, arXiv:2211.01910, 2023.

[9] J. Wei *et al.*, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 24 824–24 837.

[10] X. Wang *et al.*, "Self-Consistency Improves Chain-of-Thought Reasoning in Language Models," in *Proc. ICLR*, 2023. DOI: 10.48550/arXiv.2203.11171.

[11] D. Zhou *et al.*, "Least-to-Most Prompting Enables Complex Reasoning in Large Language Models," in *Proc. ICLR*, 2023.

[12] S. Yao *et al.*, "ReAct: Synergizing Reasoning and Acting in Language Models," in *Proc. ICLR*, 2023.

[13] S. Yao *et al.*, "Tree-of-Thoughts: Deliberate Problem Solving with Large Language Models," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[14] Z. Li *et al.*, *ReasonGraph: Visualization of Reasoning Paths*, arXiv:2503.03979, 2025.

[15] L. Cao, "GraphReason: Enhancing Reasoning Capabilities of Large Language Models Through a Graph-Based Verification Approach," in *Proc. ACL Workshop on Natural Language Reasoning and Structured Explanations*, 2024, pp. 12–24.

[16] A. Madaan *et al.*, "Self-Refine: Iterative Refinement with Self-Feedback," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[17] P. Grice, "Logic and conversation," in *Syntax and Semantics*, P. Cole and J. L. Morgan, Eds., vol. 3, New York, NY, USA: Academic Press, 1975, pp. 41–58.

[18] H. H. Clark and S. E. Brennan, "Grounding in communication," in *Perspectives on socially shared cognition*, L. B. Resnick *et al.*, Eds., Washington, DC, USA: American Psychological Association, 1991, pp. 127–149.

Machine Learning-Driven Support Algorithm for Skin Ulcers Preliminary Diagnosis: A Lightweight Approach for Digital Images Semantic Segmentation and Color-Based Classification

Debora Beneduce, Guido Pagana, Fabrizio Bertone, Giuseppe Caragnano
Fondazione LINKS – Leading Innovation and Knowledge for Society
Torino, Italy
e-mails: {debora.beneduce, guido.pagana, fabrizio.bertone, giuseppe.caragnano} @linksfoundation.com

***Abstract*—This paper presents an automated pipeline for the detection, segmentation, and severity classification of cutaneous ulcers, addressing the clinical need for objective and remote wound monitoring. Despite increasing interest, real-time and interpretable Machine Learning tools in this domain remain scarce. We propose a hybrid solution combining classical image processing and Machine Learning techniques. Exploiting the guaranteed Convolutional Neural Network performance in binary segmentation tasks, a modified U-Net architecture, trained on grayscale digital images enhanced via Contrast Limited Adaptive Histogram Equalization, achieved high segmentation performance with an Intersection over Union of 0.82, Precision of 0.93, Recall of 0.89, and Dice coefficient of 0.88, using fewer than 2 million parameters. For severity classification, superpixel-wise brightness histograms were used to extract six discriminative features. A logistic regression model trained on these features reached a classification accuracy of 94%, effectively distinguishing between ulcer classes despite intra-class variability. The system offers robust performance with fast inference of 100 milliseconds per image and skin phototype-independence.**

Keywords-machine learning; convolutional neural network; teledermatology; skin ulcers monitoring.

I. INTRODUCTION

In recent years, the increasing demand for accessible and remote healthcare services has accelerated the development of telemedicine solutions. In particular, dermatology stands out as a field where early intervention can drastically reduce long-term complications, especially in the management of chronic wounds and skin ulcers [1]. Artificial Intelligence (AI) has demonstrated remarkable progress in dermatology, particularly in the automated detection and classification of pigmented lesions and melanoma, supported by large-scale studies and Deep Learning (DL) advancements [2][3]. However, despite the clinical relevance and growing incidence of chronic wounds, the application of AI to ulcer assessment remains comparatively underexplored, with relatively few high-quality studies and limited clinical integration [4][5][6]. This imbalance highlights the need for further research on AI-driven systems tailored specifically to the complex and heterogeneous nature of cutaneous ulcers. In addition, among different kinds of skin lesions, chronic skin ulcers represent a significant clinical concern due to their prolonged healing time and resistance to standard therapeutic interventions. These lesions are defined by their failure to progress through the normal stages of wound healing, often persisting for weeks or even months. Their development is frequently associated with underlying conditions, such as immobility, diabetes, and chronic venous insufficiency, making them prevalent in elderly and at-risk populations. Clinically, chronic ulcers are commonly categorized into three primary types: pressure ulcers, diabetic foot ulcers, and venous leg ulcers [7]. Pressure ulcers–also referred to as decubitus ulcers or bedsores–are caused by sustained mechanical pressure, typically over bony prominences, which results in localized ischemia and subsequent tissue necrosis. These lesions are especially common among bedridden or immobilized individuals and are a major source of morbidity and healthcare costs [8]. Diabetic foot ulcers, on the other hand, arise due to the interplay of peripheral neuropathy, ischemia, and repeated trauma in patients with diabetes mellitus. This kind of lesion represents one of the most severe complications of diabetes and is the leading cause of non-traumatic lower-limb amputations worldwide [9]. Venous leg ulcers are primarily the result of chronic venous insufficiency, where long-term increases in venous pressure cause fluid leakage, inflammation, and eventual skin breakdown. Venous leg ulcers are the most frequently occurring type of leg ulcer and are notorious for their tendency to recur and resist conventional treatment [10].

Despite the widespread occurrence and impact of these conditions, their clinical management remains highly reliant on subjective and manual evaluations. In most cases, wound assessments are carried out through visual inspection during in-person consultations, using basic tools, such as rulers, tracing paper, or planimetry software to measure wound size. Additionally, tissue characteristics, such as color, presence of exudate, or odor, are described qualitatively, introducing high inter-observer variability and limiting the precision required for effective longitudinal monitoring [11].

In light of these challenges, there is a growing need for objective, reproducible, and accessible tools that can assist healthcare providers in the accurate evaluation and follow-up of chronic wounds. This demand is further amplified by the global shift toward telemedicine and decentralized healthcare delivery. Patients with mobility limitations or those residing in remote areas could benefit greatly from systems that enable remote wound documentation and asynchronous specialist evaluation [12]. In response to these needs, this paper presents the development of a Machine Learning (ML)-based algorithm for the real-time elaboration of digital images to detect, classify and assess the severity of cutaneous ulcers. By leveraging AI techniques—specifically Convolutional Neural Networks (CNNs)—the system performs automate key tasks,

 ISBN: 978-1-68558-312-5

such as wound detection, semantic segmentation and severity estimation based on dominant color classification.

The current study is taking part of the SALUTEDERM research project, which is dedicated to explore telemedicine techniques and solutions to enhance skin lesion healing and skin care (see Acknowledgment section for project details).

II. Related work

In the clinical field of dermatology, the diagnostic process for cutaneous lesions entails a multi-step procedure involving lesion detection, morphological assessment, and subsequent classification and staging according to established severity criteria. Visual inspection and manual palpation by clinicians remain the gold standard for ulcer classification. In some cases, tools are also used to assess lesion depth. Accurate staging is essential to determine effective treatment and reduce healing time. However, misclassifications are not uncommon, due to factors, such as skin tone variation, patient age, and overall health status [13][14]. Despite standardized assessment protocols, wound classification by visual inspection and manual palpation exhibits substantial inter-rater variability, with reported agreement coefficients ranging from poor to moderate. This variability arises from differences in clinician experience, subjective interpretation of tissue characteristics, and patient-related factors, which together challenge the reliability and reproducibility of manual ulcer staging [15]. In the last few years, to support clinical assessments, the use of digital image analysis has emerged as a promising approach to improve the evaluation of chronic wounds.

Delegating the diagnostic responsibility from human experts to automated systems involves a sequential pipeline of three fundamental and cascading tasks:

1) *Segmentation task* – automated identification and delineation of the precise boundaries of the Region of Interest (ROI) that is the ulcerated region within an image.
2) *Classification task* – division of the ROI into different classes basing on the lesion severity.
3) *Severity assessment task* – evaluation of the main characteristics extracted from the ROI.

Among the parameters useful to describe and evaluate the staging and the damage progress, to consider the kind of the involved cutaneous tissues is mandatory. In fact, the European Pressure Ulcer Advisory Panel (*EPUAP*) classified pressure ulcers into four main stages as:

- *Grade I*: intact skin with non-blanchable redness, which may also present with warmth, hardness, or pain.
- *Grade II*: partial thickness skin loss involving the epidermis and/or dermis, appearing as a shallow ulcer or blister.
- *Grade III*: full thickness skin loss extending into subcutaneous tissue often showing slough presence (yellow), but not exposing bone, tendon, or muscle.
- *Grade IV*: full thickness skin loss with extensive destruction, tissue necrosis, or damage to muscle, bone, or supporting structures [16].

Automated image analysis methods—based on bag-of-features representations and ML classifiers, such as Support Vector Machines and K-Nearest Neighbors—have been widely adopted for segmentation and classification tasks. Despite promising accuracy levels, these models often struggle when dealing with complex wound structures involving mixed tissue types, increasing the likelihood of classification errors [17][18]. Early efforts in this domain focused on basic computer vision techniques, such as color-based segmentation and edge detection, has been applied to 2D photographic data. These methods enabled semi-automated estimation of wound dimensions but were limited in their ability to capture complex tissue characteristics or to generalize across different wound types and imaging conditions [19]. The advent of ML and DL has markedly advanced the field, particularly with the introduction of CNNs for medical image analysis. CNN-based models have demonstrated strong performance in tasks, such as wound segmentation, classification of tissue types (e.g., granulation, slough, necrosis), and even prediction of healing trajectories based on sequential imaging data [20][21]. These models offer significant advantages over traditional techniques by learning hierarchical features and complex patterns directly from raw image inputs, thereby reducing the need for manual feature engineering [22].

Among the studies most closely related to the topic is the work by Zahia et al. [23] who developed a CNN-based method to classify tissue types (granulation, necrosis, and slough) using 20 high-resolution images, which were cropped into 380,000 smaller RGB patches. These were manually segmented and preprocessed with masking, grayscale conversion, Otsu thresholding, and reflection correction. Their CNN had 3 convolutional layers with increasing feature maps and used ReLU activations. They reported high sensitivity and precision for granulation and necrosis tissues (greater that 80%) but lower for slough (less than 60%). García-Zapirain et al. [24] employed a two-stage approach using a 3D CNN (DeepMedic) to extract the region of interest and segment tissues from a dataset combining original and Medtec images. Pre-processing included Gaussian smoothing and HSI color-space transformation to handle lighting variation. The first network had dual pathways for ROI detection, and the second network used four input modalities, including a prior visual appearance model built using color probability and Euclidean distances. The system achieved strong performance with Dice Similarity Coefficient (DSC) and Area Under the Curve (AUC) values around 95%. Aldughayfiq et al. [22] leveraged YOLOv5 for real-time detection and classification of pressure ulcers by grade, demonstrating a precision of 78.1%, while Pereira et al. [25] highlighted the importance of perceptually uniform color spaces (CIELAB and CIELUV) for tissue discrimination, achieving 73.8% accuracy and an AUC of 0.82. More recently, Liu et al.[4] integrated deep learning (Inception-ResNet-v2) with a clinical questionnaire in a smartphone-based diagnostic tool for pressure ulcer assessment, achieving over 90% accuracy across both cellulitis detection and necrotic tissue grading. Recent contributions have further emphasized clinical applicability, real-time performance, and quantitative wound measurement. Ramachandram et al. [26] proposed a fully automated pipeline for wound and tissue segmentation on mobile devices, employing two CNNs to segment the

ulcer and classify tissue types, such as epithelial, granulation, slough, and eschar. Their models achieved an Intersection over Union (IoU) of 0.8644 for wounds and 0.7192 for tissue classification, while also quantifying the substantial inter- and intra-rater variability among clinicians. Liu et al. [27] combined U-Net and Mask R-CNN with LiDAR-based area measurement for pressure injuries, reporting a Dice coefficient of 0.8448 on external validation and a mean relative area error of 26.2% compared to manual measurements, highlighting both the potential and the current limitations of quantitative wound assessment in clinical settings. Carvalho et al. [28] explored CNN and Transformer-based architectures, including DeepLabV3+, SegFormer, and MedSAM, for segmentation and real-world wound measurement. Using reference markers to scale images, their pipeline achieved Dice scores above 92% and area estimation errors as low as 5.36% on private datasets, although performance decreased when the entire pipeline was applied under diverse imaging conditions.

Despite these advances, several challenges persist. Learning-based methods require large volumes of high-quality annotated medical images, which remain scarce due to high annotation costs, limited patient data, and ethical constraints [29]. Unlike imaging modalities such as brain, retinal, or chest CT, dermatological conditions have traditionally been assessed via direct visual inspection, complicating the creation of large-scale datasets. Additional challenges include differentiating among tissue types, dealing with ill-defined lesion boundaries, and ensuring robustness to variations in lighting, skin tone, and image quality [30][31]. Data augmentation strategies have been proposed to mitigate small dataset limitations [22][32]. Collectively, these studies indicate that automated segmentation, tissue classification, and measurement are promising, but fully integrated systems capable of robust performance across heterogeneous conditions, severity assessment, and telemedicine deployment remain an open research need.

III. METHODS

Building on the presented premises, the present study aims to develop an algorithm that first achieves high-quality semantic segmentation of skin ulcers, and subsequently enables color-based lesion severity evaluation, with limited computational demands. Indeed, achieving strong segmentation accuracy is crucial, but it must be balanced with low computational cost–implying a model with fewer learnable parameters and lightweight architecture–to ensure real-world efficiency [33]. For this reason, small network sizes and high inference speeds guaranteed by CNNs architectures have been preferred over superior segmentation accuracy performed by more complex methods, like transformer-based ones [34].

The proposed algorithm is structured as a sequential pipeline, in which each task takes as input the output of the previous one. Specifically, the process begins with a pre-processing stage to standardize and enhance the input data, followed by lesion segmentation with associated post-processing to refine the obtained masks, and finally lesion classification based on the segmented regions. This design ensures modularity, efficiency, and a clear propagation of information across tasks.

The entire system has been programmed in Python language and built, trained, validated and tested on two public databases from the 2021 MICCAI Foot Ulcers Segmentation Challenge, the first composed by over 1200 de-identified diabetic foot ulcers images and their respective labels [35], and the second created in collaboration with the AZH Wound & Vascular Center composed by 1109 cropped patches of foot ulcers [36].

A. Segmentation task

The first step of the pipeline is the ROI detection that consists in the recognition of the ulcer area and its borders highlighting.

Pre-processing: To perform the task, the dataset variability reduction is recommended as a starting pre-processing strategy. For this reason the images from the dataset need to be normalized both in terms of size and in terms of pixels values to make the training consistent and stable. Hence, a division by 255 – the maximum pixel value– is applied to each of the three color channels of the RGB color space, namely Red, Green and Blue, so that the operational range per pixel is now $[0, 1]$; then each image undergoes a `resize` operation to the standard 512×512 size. The size choice is a trade-off between the need to preserve as much information as possible and a reduced computational weight. It has also been demonstrated that the exclusion of the brightness information, which carries a big amount of variability and translates into the use of single-channel grayscale (GS) images instead of three-channels RGB ones, can further improve segmentation performance in dermatological studies by making the algorithm less sensitive to illumination artifacts [37]. For this reason, a comparison between ulcers detection on RGB and on GS images, both original and contrast-enhanced, will be evaluated.

Training: Among the most efficient and cost-effective neural network architecture employed for binary segmentation, U-Net demonstrates relevant potential. We implemented a deep U-Net variant tailored for high-resolution biomedical image segmentation tasks. The network adopts a symmetric encoder–decoder architecture with four levels of downsampling and upsampling, and includes skip connections to preserve spatial context and fine-grained features. The encoder consists of four convolutional blocks, each composed of two convolutional layers (kernel size 3×3, ReLU activation, He-normal initialization, and same padding), followed by 2×2 max pooling for downsampling. The number of filters doubles with each level, starting from 16 up to 128. A bottleneck layer with two convolutional layers and 256 filters processes the compressed representation. The decoder mirrors the encoder structure. Each upsampling step (via 32×2 upsampling) is followed by a concatenation with the corresponding encoder feature map (skip connection), and two convolutional layers that progressively reduce the number of filters back to 16. The final layer is a 1×1 convolution with sigmoid activation, producing a single-channel probability map for binary segmentation. This architecture balances depth and computational efficiency, maintaining the U-Net's ability to integrate multi-scale contextual information while enabling the extraction of

deeper features via the added encoder stage. The modified U-Net architecture proposed by the authors is depicted in Figure 1 and more details, as the number of layers, are reported for clarity in Table I.

TABLE I. ARCHITECTURE OF OUR U-NET-LIKE MODEL.

Layer(s) block	Feature Maps	Kernel Size
Input	1	-
Encoder 1	16 → 16	3×3, 3×3
Pooling 1	16	2×2 (max pooling)
Encoder 2	32 → 32	3×3, 3×3
Pooling 2	32	2×2 (max pooling)
Encoder 3	64 → 64	3×3, 3×3
Pooling 3	64	2×2 (max pooling)
Encoder 4	128 → 128	3×3, 3×3
Pooling 4	128	2×2 (max pooling)
Bottleneck	256 → 256	3×3, 3×3
Decoder 1	256+128 → 128 → 128	up 2×2, 3×3, 3×3
Decoder 2	128+64 → 64 → 64	up 2×2, 3×3, 3×3
Decoder 3	64+32 → 32 → 32	up 2×2, 3×3, 3×3
Decoder 4	32+16 → 16 → 16	up 2×2, 3×3, 3×3
Output	1	1×1 (sigmoid)

Training parameters, as the number of training epochs, batch size and learning rate, and data augmentation techniques have been tuned after several trainings. The training has been run across 50 epochs with $batch\_size = 2$ and $learning\_rate = 1e^{-4}$. The Adam optimizer has been set to take advantage of:

- Automatically regulated learning rate useful to manage ulcers borders that can produce different gradients in respect to other areas.
- Faster convergence in presence of a U-Net architecture.
- Low sensitivity to unbalanced classes (e.g. 90% background, 10% object).

Data augmentation techniques included clockwise/counterclockwise rotations, width/height shifts, zoom and horizontal/vertical flips and the fill mode was set as `'nearest'`. Due to this method, the volume of data has increased fourfold. Finally, the 80% of the whole dataset was split using a 90/10 ratio between training and validation sets and both RGB and grayscale enhanced images have been used for comparison in different trainings.

Unlike many previous studies, where ulcer analysis is performed in two sequential steps—first identifying a Region Of Interest (ROI) and then segmenting the wound within that ROI—we directly trained the U-Net to segment the ulcer from the entire image. This decision was motivated by both methodological and practical considerations. First, ROI detection introduces an additional preprocessing stage that may propagate errors and increase variability across images. Second, U-Net architectures have proven effective at simultaneously learning global contextual cues and local boundary information, allowing reliable segmentation even without prior cropping [38]. Finally, an end-to-end segmentation pipeline reduces complexity and enhances reproducibility, making the method easier to deploy in real-world clinical workflows.

Post-processing: The predicted binary masks generated by the algorithm on the test set (remaining 20% of the dataset) have been observed and analyzed. The biggest issue came from the background noise represented by disturbing elements present into the image and characterized by a range of colors similar to the one of the skin ulcers. Also, very small areas of healthy skin are sometimes wrongly identified as ROIs. After a `resize` operation at the original size of each input image, these considerations led to two main post-processing procedures:

1) Removal of very small object–For each mask, each segmented object with $area < area_{max}/6$, where $area_{max}$ is the area (expressed in pixels) of the biggest recognized object, is ignored.
2) Background removal through skin segmentation–In addition to different appearances of ulcers of different etiology and class, healthcare professionals usually take pictures in diverse ways, in heterogeneous light conditions and position. These factors further increase the variability of the dataset, which is already limited in size given the complexity of the problem. To limit confounding elements and reduce the amount of data to be analyzed, a strategy to isolate the affected limb by removing the background has been implemented. First, images are converted from *RGB* color-space to YC_rC_b one, useful to take advantage of the separation between the luminance information represented by the *Y* channel and the chrominance contribution expressed by the C_i channel, where subscript *i* can stand for red (C_r) or blue (C_b). The luminance represents the brightness level of the image, whereas the blue and red chrominances carry the color information by indicating the shift of blue and red channels from the luminance value [39]. As visible in Figure 2, the C_r channel shows the cutaneous area highlighted in respect to other objects into the image offering the strategy for skin segmentation.

B. Classification task

Among the various features used for skin chronic wound classification, color information plays a prominent role due to its partial correlation with the depth and extent of tissue damage [40]. Indeed, first stage lesions, partial tissue loss lesions, covered by slough lesions and necrotic lesions are characterized by colors ranging from light red to vivid red to yellow-tinged to brown/black. As the available dataset is made of diabetic foot ulcers, Stage I lesions are not represented and therefore are not recognized during the segmentation task. Since the classification module operates exclusively on segmented ulcers, the absence of Stage I lesions in segmentation directly implies that their classification is not included in the scope of the algorithm.

To enable chromatic differentiation of lesions, brightness histograms of each R, G and B channel intensities were computed and analyzed. The histograms show the possible intensity values, $[0, 255]$ along the x-axis, and the frequency of pixels exhibiting each intensity value along the y-axis. Figure 3 displays an example of the different distributions of colors intensities among different stages of ulcers.

To reduce the huge amount of data deriving from the

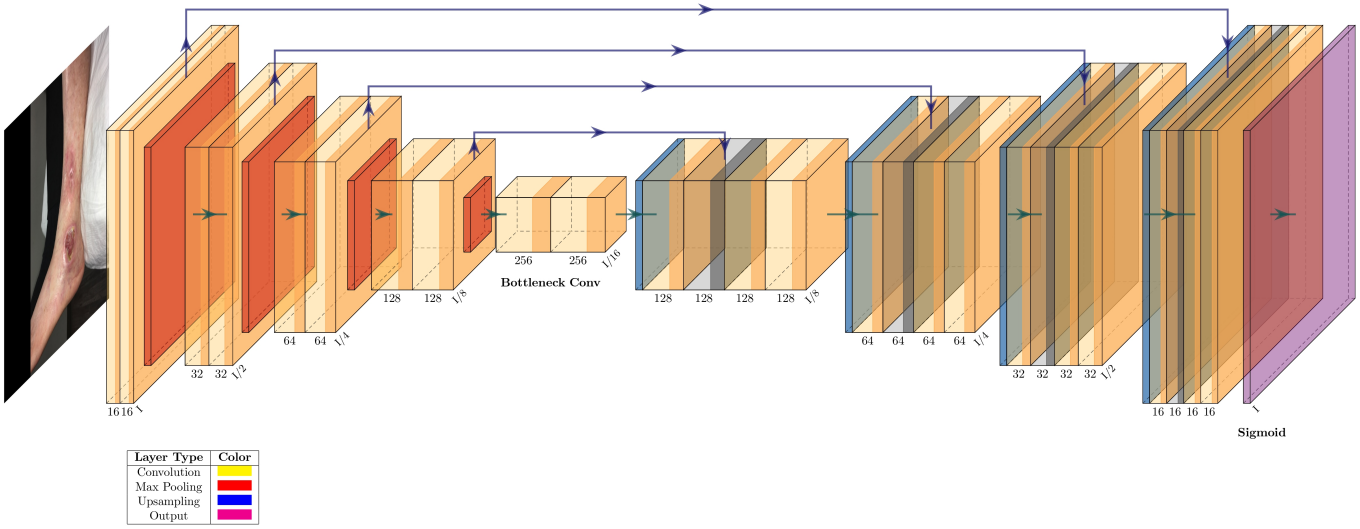


Figure 1. Modified U-Net architecture proposed by the authors.

Figure 2. RGB image [17], YC_rC_b image, Y channel, C_r channel, and C_b channel (from left to right).

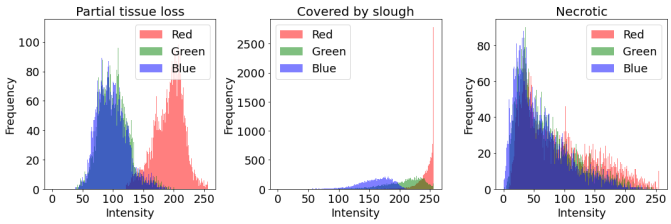


Figure 3. Partial tissue loss ulcer, necrotic ulcer and covered by slough ulcer and respective brightness R, G and B histograms (from left to right).

processing of each single pixel, the related introduced variability and because regions of pixels are more informative than single pixels, the Simple Linear Iterative Clustering (SLIC) algorithm–based on *Superpixel* technique–is performed. This method segments an image into a chosen number of regions–named superpixels– clustering pixels that appear similar according to some perceptual features relying on measures based not only on color similarity but also on the shape of the regions delimiting areas significant changes in intensity [41].

As the dataset does not provide any class or tissue information, the involvement of a medical expert for accurate labeling was crucial. Consequently, the entire database was carefully reviewed by the clinician, who selected 216 lesions deemed representative of all severity classes, while deliberately avoiding borderline cases that could introduce ambiguity or bias due to subjective interpretation. Each selected lesion was then manually classified under the expert's supervision, ensuring that the labels reflected both clinical relevance and consistency. From each of these lesions, six parameters has been extracted and, in addiction to the ground truth labels, fed into a multiple logistic regressor for the class prediction by splitting the set of labeled ulcers into 80% training set and 20% test set.

In the following section, the results achieved through the implemented workflow are presented.

IV. Results Analysis

The results are organized according to the pipeline introduced in Section II and Section III.

A. Segmentation task results

The described modified U-Net model has been trained across 50 epochs. A notably small $batch\_size = 2$ yielded more favorable training dynamics compared to larger batch sizes. The increased stochasticity in the gradient estimates, induced by the smaller batch size, likely acted as an implicit regularization mechanism, contributing to improved generalization and reduced overfitting also lowering GPU memory usage. The most performing trainings derived from GS images training, in particular on locally-contrast modified GS images via CLAHE method. Contrast Limited Adaptive Histogram Equalization (CLAHE) is an image processing technique that enhances the contrast of images by applying histogram equalization locally, in small regions, rather than globally. CLAHE also incorporates a contrast-limiting step to prevent over-amplification of noise in homogeneous regions. It does this by clipping the histogram at a predefined threshold ($clip\_limit$ parameter) before redistributing the clipped pixels evenly. After several trials, the optimal value has been found to be $clip\_limit = 0.8$. An example of comparison among *RGB*, *GS* and *GS*-contrast enhanced outcomes is reported in Figure 4.

Despite the discrete performance achieved, the segmentation of confounding elements remained a challenge. For this reason, post-processing is compulsory both in terms of small object removal and in color-thresholding to discard out-of-skin segmented element. Since in some cases the lesion is extremely small, while in others similarly sized objects

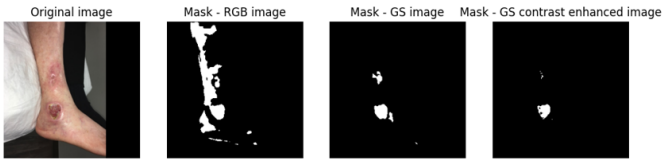


Figure 4. Example of comparison on a test image [35] between binary segmentation masks predicted by training on RGB, GS and GS-contrast enhanced images.

represent confounding elements, it is not feasible to define a global size threshold that is consistently valid across all images. To overcome this problem, it has been observed that the biggest recognized object is always represented by the real lesion to be segmented; in respect to this object, other selected areas, smaller than $\frac{1}{6}$ of the biggest identified area on the same image, are for sure segmentation errors. In case only one object is found, then it is considered as an ulcer unless it is discovered not to belong to the skin area detected through the second post-processing step. As illustrated in Section III, the YC_rC_b color-space can be helpful for skin segmentation. Indeed, by executing a global thresholding in the range of $[0.55, 0.70]$ for C_r channel, it is possible to easily remove the background (Figure 5). Each detected object outside the segmented skin is then ignored.

(a) (b)

Figure 5. Skin segmentation: (a) original image [35], (b) result after global thresholding on channel C_r.

At the end, the final best training provided good segmentation performance represented by $IoU = 0.82, Precision = 0.93, Recall = 0.89$ and $Dice\ Coefficient = 0.88$. Being in a biomedical field, the Precision metric gains more importance in respect to the Recall one since to reduce the amount of false negative pixels (in this case meaning non-detected ulcers) is considered more relevant than to have less false positive pixels (meaning healthy skin recognized as damaged). Thus, the elevated precision score is of particular relevance in study scenarios where accurate discrimination is as inherently difficult as relevant. Finally, the mean inference time per image is 100 ms.

It is important to emphasize that our modified U-Net achieves competitive results while maintaining a notably low number of trainable parameters — fewer than 2 million. Table II presents *IoU* scores reported by some of the most frequently cited studies within the same research domain, employing comparable ML-based segmentation methodologies, alongside their respective parameter counts. To ensure a fair and meaningful evaluation, we compared our proposed model against several widely used and high-performing U-Net variants in biomedical image segmentation, namely the standard U-Net, ResU-Net, Attention U-Net, and FU-SegNet. These architectures represent the most commonly adopted baselines in the field and cover different directions of improvement to the original U-Net design, such as residual connections, attention mechanisms, and enhanced decoder structures. This choice allows us to directly assess the specific contribution of our modifications within the same architectural family.

Finally, three examples are displayed in Figure 6 to show the segmentation outcomes. The first image also assesses the phototype-independence of the model, which should not be taken for granted. Although the dataset does not provide explicit information on skin phototype, a visual inspection of the database revealed that only about 5% of the images correspond to darker skin tones. Nevertheless, the first image on the left demonstrates that, despite this underrepresentation, darker phototypes do not compromise the robustness of the algorithm.

TABLE II. COMPARISON WITH STATE-OF-THE-ART ULCER SEGMENTATION RESEARCHES.

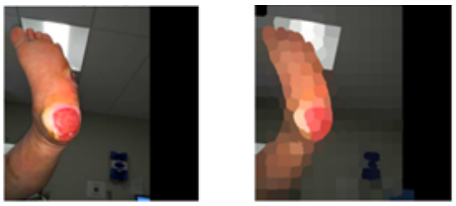
Model	IoU	Param.	Reference
Standard U-Net	0.68	7.8M	[38]
ResU-Net	0.72	8.9M	[42]
Attention U-Net	0.75	8.9M	[43]
FUSegNet	0.77	12.7M	[32]
Authors' Modified U-Net	0.82	1.9M	–

Figure 6. Examples of segmentation outcomes on test images [35].

B. Classification task results

The brightness histograms approach allows for the extraction of color distribution patterns within each lesion, facilitating the identification of relevant visual features, such as variations in redness, yellowness, or darkness associated with different tissue types or stages of wound healing. The use of this method would be computationally and time-consuming if applied pixel by pixel. For this reason, the SLIC algorithm paired with histograms analysis is the key of our solution to the classification problem. Considering *N* as the number of pixels in the input image, *K* as the desired number of superpixels, the approximate size of a superpixel will be N/K and for roughly equally sized superpixels there would be a center at every grid interval $S = \sqrt{\frac{N}{K}}$, resulting into a superpixel spatial extent about S^2. The input parameter for the SLIC algorithm are then the input image, the parameter *K* and the variable **m** that is a measure of superpixel compactness. For the current study, $K = 290$ and $m = 20$ have been demonstrated to guarantee the best tradeoff between minimizing color-variability and computational costs and avoiding the loss of original color

and shape information (Figure 7).



(a) (b)

Figure 7. SLIC technique: (a) original image [35], (b) SLIC processed image with $K = 290$ and $m = 20$.

Each superpixel has then been enumerated so that the interested areas could be manually selected by typing their respective number. RGB histograms were computed for each extracted superpixel, and six parameters–identified as the most discriminative–were evaluated. These consist of the median values of the three color channels and the intensity differences among them. The ground-truth labels and the respective six variables for each lesion have been exploited to train a multiple logistic regressor. The extracted variables for the training set are reported in Figure 8 to have an overview of the distributions. The total number of samples is higher than 216 because some lesion are mapped by more than a single superpixel. In addition, some samples from healthy skin areas has been analyzed and inserted into the figure to demonstrate, once more, how the distinction between injured and non-injured skin is not trivial. Indeed, the extracted parameters from healthy skin result into overlapping to all of the three ulcers classes. Moreover, the limited representation of the necrotic lesion class can be observed. This problem is counterbalanced by the evident color separation from the other classes.

Despite the quite considerable variability of the features given by different reasons (e.g. image acquisition device and the huge variety of ulcers) the extracted features demonstrated to be sufficiently discriminatory. Figure 9 present the data as mean ± Standard Error of the Mean (SEM) in order to provide a concise summary of class-level differences. While the SEM does not reflect the within-class variability, it effectively represents the precision of the estimated class means, allowing for clearer visualization of systematic differences between classes. This choice is further supported by the model's high classification accuracy of 94%, suggesting that, despite the high variability, the method proves to be both effective and reliable. The notable inter-class discrimination ability is, furthermore, expressed by the confusion matrix in Figure 10. The vertical axis represent the ground truth class, whereas the horizontal axis represents the class predicted by our model.

Classes are assessed according to the *EPUAP* definition and, as already discussed, class I is not considered as it is not represented by the dataset. The classification error reaches only 6% in the higher severity classes. Expanding the dataset with a larger number of labeled images would enable a more robust evaluation, potentially confirming that the majority of misclassifications are conservative–i.e., the predicted severity tends to be higher than the actual one–an outcome that is generally preferable in medical contexts to avoid underestimation of critical conditions.

With regard to the propagation of segmentation errors to the sequential classification task, although the segmentation model does not achieve 100% accuracy in terms of IoU, this does not cause a meaningful drop of performance. Segmentation inaccuracies mainly occur at the lesion borders, whereas in clinical practice the most severely affected tissue is often located in the central region of the ulcer — a region that tends to heal more slowly, and which becomes the critical determinant of severity. Consequently, since that central area is generally well captured by the model, the predicted class remains accurate and robust despite minor segmentation errors.

V. Conclusion and Future work

This work presented the design and evaluation of an end-to-end system for the automated analysis of cutaneous ulcers, addressing two critical tasks in the wound care pipeline: segmentation and severity classification. The proposed solution leverages a combination of classical image enhancement techniques, DL architectures, and lightweight ML classifiers to support clinicians in the timely assessment and monitoring of chronic wounds.

For the segmentation task, a modified version of the U-Net architecture was employed. The model, trained on grayscale images enhanced via the CLAHE algorithm, demonstrated superior performance compared to RGB-based approaches. CLAHE proved particularly beneficial in enhancing local contrast while avoiding noise over-amplification, allowing for more reliable lesion boundary detection. The model achieved a mean Intersection over Union (IoU) of 0.82, Precision of 0.93, Recall of 0.89, and a Dice coefficient of 0.88, which compare favorably to existing state-of-the-art solutions while maintaining a significantly lower number of parameters (< 2M). This makes the proposed model suitable for real-time deployment, especially in resource-constrained clinical or mobile environments. Despite the strong performance, certain challenges were observed in the segmentation of confounding elements, such as artifacts or visually similar skin regions. These were effectively addressed through a post-processing pipeline, which included object size filtering and color-based thresholding in the YC_rC_b color space. In particular, lesions were robustly distinguished from non-skin elements by analyzing the C_r channel, enabling the removal of out-of-context segmented areas.

For the classification task, the model relied on the extraction of color-based features from wound superpixels obtained using the SLIC algorithm. This method significantly reduced the computational burden compared to pixel-level analysis, while preserving the spatial and chromatic properties of the lesions. The most informative features–channel medians and inter-channel differences–were used to train a multiple logistic regression classifier, which achieved an overall classification accuracy of 94%. Notably, the model performed well despite significant intra-class variability due to differences in ulcer morphology, acquisition conditions, and lighting. The use of mean ± SEM to present class-level feature distributions

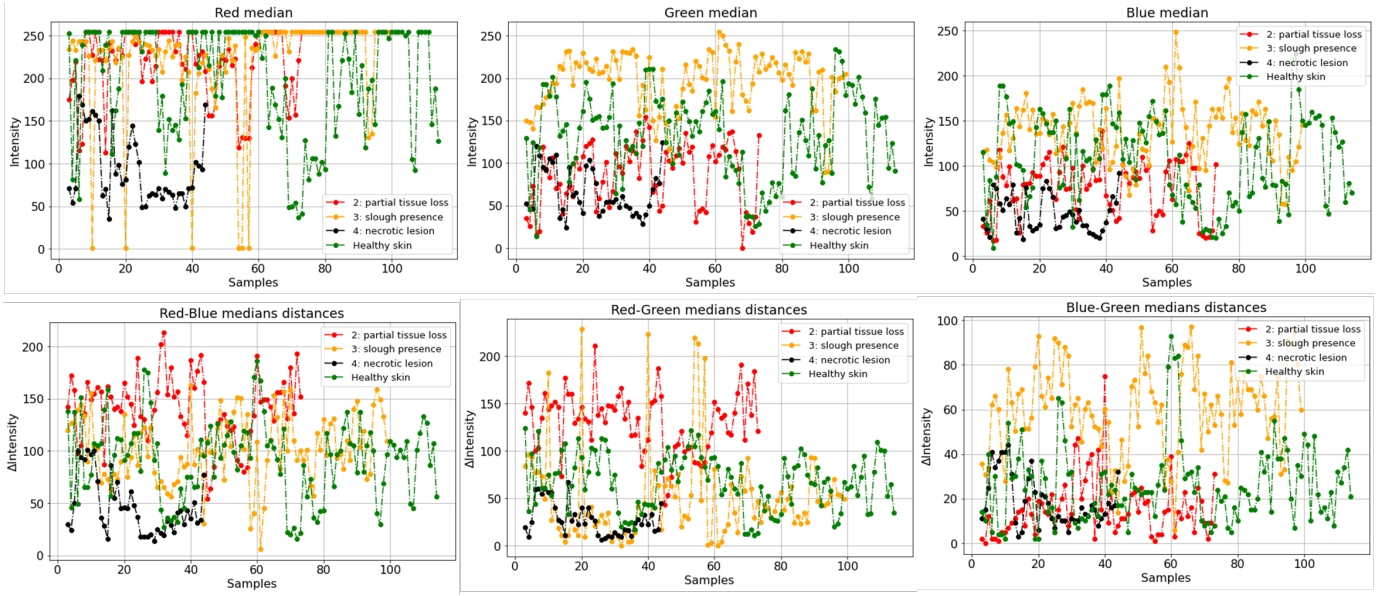


Figure 8. Extracted features for each class (training set): Red, Green and Blue medians (first raw, from left to right); Red-Blue, Red-Green and Blue-Green medians distances (second raw, from left to right).

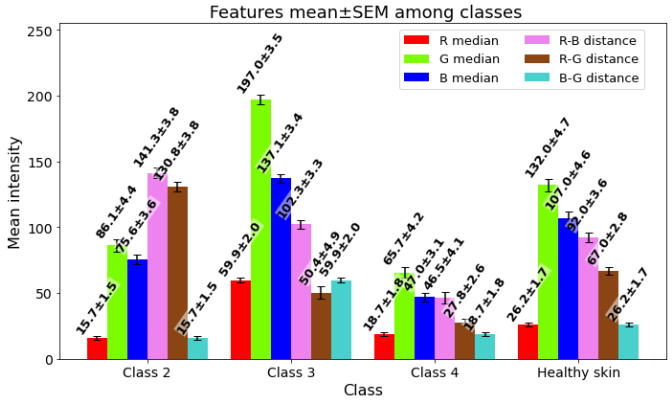


Figure 9. Mean of the extracted features for classification task.

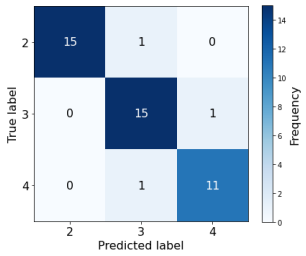


Figure 10. Confusion matrix for the classification task.

proved to be an effective strategy for visually conveying discriminative trends, even in the presence of overlapping data. Importantly, the classification error was predominantly observed in the higher severity classes, where conservative misclassifications are preferable in clinical settings, as they reduce the risk of underestimating potentially critical conditions. Moreover, the model demonstrated promising phototype-independence, which is a crucial factor for broad applicability in diverse patient populations.

Building upon the encouraging results obtained in both segmentation and classification tasks, future developments will focus on three main directions. First, authors plan to significantly expand and diversify the dataset, with particular emphasis on including underrepresented ulcer classes (such as EPUAP Stage I and various necrotic subtypes) and a broader range of patient skin phototypes. This will improve the generalizability of the model and ensure its reliability in real-world, heterogeneous clinical settings. Second, efforts will be directed toward extracting clinically interpretable outcome measures from the segmentation and classification results. These outcomes will be used to derive a quantitative severity score for each lesion, analogous to established metrics in the literature—such as the Photographic Wound Assessment Tool (PWAT)—but designed to be computationally lighter and fully automatable, thus more suited for integration into digital health systems. Third, all modules of the proposed pipeline will be integrated into a dedicated telemedicine device, currently under development by the authors. This device will be capable of acquiring standardized digital images of the wound, will incorporate the proposed algorithms to provide preliminary, automated assessments of wound presence and will evaluate the temporal evolution of the severity score.

The system has been trained, validated, and tested on anonymized images from a public dataset that already complies with General Data Protection Regulation (GDPR) requirements. Regarding its future clinical use, procedures for approval by the local ethics committee have already been initiated, and a dedicated protocol for data encryption and anonymization will be developed to ensure secure storage in the databases that will be progressively built.

The system is designed to function as a medical decision-support tool, especially in settings where specialist access is limited, such as home care or remote rural areas. This integrated approach–combining robust AI algorithms, clinically

meaningful outcomes, and practical hardware implementation–lays the foundation for a comprehensive and scalable solution in the emerging field of AI-assisted wound telemonitoring.

ACKNOWLEDGMENT

This research has been funded under the "Aggregazioni R&S – Salute" call issued by Assessorato Sviluppo Economico, Formazione e Lavoro, Trasporti e Mobilità Sostenibile, Ricerca, Innovazione e Trasferimento Tecnologico of the Aosta Valley Region, with the project SALUTEDERM (CUP B49J23008170009).

REFERENCES

[1] E. L. Eber, E. Arzberger, C. Michor, R. HofmannWellenhof, and W. Salmhofer, *Mobile Teledermatologie in der Behandlung chronischer Ulzera*, German, 2019. DOI: 10 . 1007 / S00105-019-4397-5.

[2] N. Melarkode, K. Srinivasan, S. M. Qaisar, and P. Plawiak, "AI-Powered Diagnosis of Skin Cancer: A Contemporary Review, Open Challenges and Future Research Directions", *Cancers*, vol. 15, no. 4, p. 1183, Feb. 2023, Epub 2023 Feb 13. DOI: 10.3390/cancers15041183.

[3] C. Lei et al., "Convolutional Neural Network Models for Visual Classification of Pressure Ulcer Stages Cross-Sectional Study", *JMIR Medical Informatics*, vol. 13, no. 1, Mar. 2025. DOI: 10.2196/62774.

[4] T. J. Liu et al., "A pressure ulcers assessment system for diagnosis and decision making using convolutional neural networks", *Journal of the Formosan Medical Association*, vol. 121, no. 11, pp. 2227–2236, Nov. 2022, Epub 2022 May 4, ISSN: 0929-6646. DOI: 10.1016/j.jfma.2022.04.010.

[5] P.-H. Huang et al., "Development of a deep learning–based tool to assist wound classification", *Journal of Plastic, Reconstructive & Aesthetic Surgery*, vol. 76, no. 6, pp. 1462–1470, Jun. 2023, Epub 2023 Feb 10. DOI: 10.1016/j.bjps.2023.01.015.

[6] R. Niri et al., "Wound Segmentation with UNet Using a Dual Attention Mechanism and Transfer Learning", *Journal of Imaging Informatics in Medicine*, Jan. 2025, Published 23 Jan 2025. DOI: 10.1007/s10278-025-01386-w.

[7] R. G. Frykberg and J. Banks, "Challenges in the Treatment of Chronic Wounds", *Advances in Wound Care*, vol. 4, no. 9, pp. 560–582, Sep. 2015. DOI: 10.1089/wound.2015.0635.

[8] L. E. Edsberg et al., "Revised National Pressure Ulcer Advisory Panel Pressure Injury Staging System", *Journal of Wound, Ostomy and Continence Nursing*, vol. 43, pp. 585–597, 6 Nov. 2016, ISSN: 10715754. DOI: 10 . 1097 / WON . 0000000000000281.

[9] D. G. Armstrong, A. J. Boulton, and S. A. Bus, "Diabetic Foot Ulcers and Their Recurrence", *New England Journal of Medicine*, vol. 376, pp. 2367–2375, 24 Jun. 2017, ISSN: 0028-4793. DOI: 10.1056/nejmra1615439.

[10] T. F. O'Donnell et al., "Management of venous leg ulcers: Clinical practice guidelines of the Society for Vascular Surgery® and the American Venous Forum", *Journal of Vascular Surgery*, vol. 60, 3S–59S, 2 2014, ISSN: 10976809. DOI: 10.1016/j.jvs.2014.04.049.

[11] K. Ousey, B. Gilchrist, and H. Jaimes, "Understanding clinical practice challenges: a survey performed with wound care clinicians to explore wound assessment frameworks", *Wounds International*, vol. 9, no. 4, pp. 58–62, 2018, ©Wounds International 2018, [retrieved: August, 2025].

[12] C. M. Contreras et al., "Telemedicine: Patient-Provider Clinical Engagement During the COVID-19 Pandemic and Beyond", *Journal of Gastrointestinal Surgery*, vol. 24, no. 8, pp. 1692–1697, 2020. DOI: 10.1007/s11605-020-04623-5.

[13] M. C. Araujo, A. R. Silva, and F. A. Pereira, "A Systematic Review on Wound Classification Challenges in Clinical Practice", *Journal of Wound Care*, vol. 30, no. 3, pp. 132–140, 2021.

[14] M. Cabrera, E. Gómez, and A. Torres, "Pressure Injury Image Analysis with Machine Learning Techniques: A Systematic Review", *Journal of Biomedical Informatics*, vol. 107, p. 103 432, 2020.

[15] R. G. Sibbald, D. Krasner, and J. Lutz, "Challenges in wound assessment: Interrater variability and implications for clinical care", *Advances in Skin & Wound Care*, vol. 32, no. 8, pp. 388–395, 2019, [retrieved: July, 2025]. DOI: 10 . 1097 / 01 . ASW . 0000565797.24804.52.

[16] National Pressure Ulcer Advisory Panel and European Pressure Ulcer Advisory Panel, "Prevention and Treatment of Pressure Ulcers/Injuries: Clinical Practice Guideline", International Guideline Development Group, Cambridge Media, Osborne Park, Western Australia, Tech. Rep., 2019, Also developed in collaboration with Pan Pacific Pressure Injury Alliance; third edition released 15 Nov 2019.

[17] L. Wang, X. Zhou, and H. Zhang, "Deep LearningBased Classification of Pressure Injuries Using Multimodal Imaging", *Theranostics*, vol. 15, no. 7, pp. 1662–1675, 2023.

[18] Y. Jiang, P. Liu, and J. Wang, "Automated Wound Classification Using Machine Learning Algorithms", *Computers in Biology and Medicine*, vol. 127, p. 104 057, 2020.

[19] L. B. Jørgensen, J. A. Sørensen, G. B. Jemec, and K. B. Yderstræde, "Methods to assess area and volume of wounds – a systematic review", *International Wound Journal*, vol. 13, pp. 540–553, 4 Aug. 2016, ISSN: 1742481X. DOI: 10.1111/iwj.12472.

[20] D. M. Anisuzzaman et al., "Image Based Artificial Intelligence in Wound Assessment: A Systematic Review", *Adv Wound Care (New Rochelle)*, vol. 11, pp. 687–709, 12 Dec. 2022, Epub 2021 Dec 20; published Sep 21 2021. DOI: 10.1089/wound.2021.0091.

[21] F. Veredas, H. Mesa, and L. Morente, "Binary tissue classification on wound images with neural networks and bayesian classifiers", *IEEE Transactions on Medical Imaging*, vol. 29, pp. 410–427, 2 Feb. 2010, ISSN: 02780062. DOI: 10.1109/TMI.2009.2033595.

[22] B. Aldughayfiq, F. Ashfaq, N. Z. Jhanjhi, and M. Humayun, "YOLO-Based Deep Learning Model for Pressure Ulcer Detection and Classification", *Healthcare*, vol. 11, no. 9, p. 1222, 2023, Received: 12 Mar 2023; Revised: 15 Apr 2023; Accepted: 22 Apr 2023; Published: 25 Apr 2023. DOI: 10.3390/healthcare11091222.

[23] S. Zahia, D. Sierra-Sosa, B. Garcia-Zapirain, and A. Elmaghraby, "Tissue classification and segmentation of pressure injuries using convolutional neural networks", *Computer Methods and Programs in Biomedicine*, vol. 159, pp. 51–58, Jun. 2018, Epub 2018 Mar 3, ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2018.02.018.

[24] B. García-Zapirain, M. Elmogy, A. El-Baz, and A. S. Elmaghraby, "Classification of pressure ulcer tissues with 3D convolutional neural network", *Medical & Biological Engineering & Computing*, vol. 56, no. 12, pp. 2245–2258, Dec. 2018, Epub 2018 Jun 15, ISSN: 0140-0118. DOI: 10.1007/s11517-018-1835-y.

[25] S. M. Pereira, M. A. Frade, R. M. Rangayyan, and P. M. A. Marques, "Classification of color images of dermatological ulcers", *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 1, pp. 136–142, Jan. 2013, Epub 2012 Nov 15, ISSN: 2168-2194. DOI: 10.1109/TITB.2012.2227493.

[26] D. Ramachandram et al., "Fully Automated Wound Tissue Segmentation Using Deep Learning on Mobile Devices", *JMIR mHealth and uHealth*, vol. 10, e36977, 2022. DOI: 10.2196/36977.

[27] T. J. Liu et al., "Automatic Segmentation and Measurement of Pressure Injuries Using Deep Learning Models and a LiDAR Camera", *Scientific Reports*, vol. 13, p. 680, 2023. DOI: 10.1038/s41598-022-26812-9.

[28] D. Carvalho et al., "Enhancing Chronic Wound Assessment through Agreement Analysis and Deep Learning Models", *Scientific Reports*, vol. 15, p. 12 345, 2025. DOI: 10.1038/s41598-025-06703-5.

[29] N. Tajbakhsh et al., "Embracing Imperfect Datasets: A Review of Deep Learning Solutions for Medical Image Segmentation", *Medical Image Analysis*, vol. 63, p. 101 693, 2020. DOI: 10.1016/j.media.2020.101693.

[30] M. A. Kabir, N. Roy, M. E. Hossain, J. Featherston, and S. Ahmed, *Deep Learning for Wound Tissue Segmentation: A Comprehensive Evaluation using A Novel Dataset*, [retrieved: July, 2025], 2025. arXiv: 2502.10652 [eess.IV].

[31] H. Liu et al., "Current status, challenges, and prospects of artificial intelligence applications in wound repair theranostics", *Theranostics*, vol. 15, pp. 1662–1688, 2025, [retrieved: June, 2025]. DOI: 10.7150/thno.105109.

[32] Y. Patel et al., "Integrated image and location analysis for wound classification: a deep learning approach", *Scientific Reports*, vol. 14, 1 2024, ISSN: 20452322. DOI: 10.1038/s41598-024-56626-w.

[33] C. BroniBediako, J. Xia, and N. Yokoya, "RealTime Semantic Segmentation: A Brief Survey & Comparative Study in Remote Sensing", *IEEE Geoscience and Remote Sensing Magazine*, pp. 2–33, 2023. DOI: 10.1109/MGRS.2023.3321258.

[34] J. Cheng, H. Li, D. Li, S. Hua, and V. S. Sheng, "A Survey on Image Semantic Segmentation Using Deep Learning Techniques", *Computers, Materials & Continua*, vol. 74, no. 1, pp. 1941–1957, 2023. DOI: 10.32604/cmc.2023.032757.

[35] C. Wang et al., *FUSeg: The Foot Ulcer Segmentation Challenge Dataset*, GitHub repository, [retrieved: May, 2025], 2021.

[36] D. M. Anisuzzaman, Y. Patel, J. Niezgoda, S. Gopalakrishnan, and Z. Yu, *AZH Wound Care Center Dataset (patches)*, GitHub repository, Preprocessed wound image patches from clinical data collected at the AZH Wound and Vascular Center, Milwaukee, WI, [retrieved: May, 2025], 2020.

[37] Y. N. Hwang, M. J. Seo, and S. M. Kim, "A Segmentation of Melanocytic Skin Lesions in Dermoscopic and Standard Images Using a Hybrid TwoStage Approach", *BioMed Research International*, vol. 2021, p. 5 562 801, 2021. DOI: 10.1155/2021/5562801.

[38] O. Ronneberger, P. Fischer, and T. Brox, "UNet: Convolutional networks for biomedical image segmentation", in *Medical Image Computing and ComputerAssisted Intervention – MICCAI 2015*, Springer, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

[39] International Telecommunication Union, *Recommendation ITUR BT.601-7: Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios*, https://www.itu.int/rec/R-REC-BT.601, Accessed: 2025-07-01, 2011.

[40] S. Li, A. H. Mohamedi, J. Senkowsky, A. Nair, and L. Tang, "Imaging in Chronic Wound Diagnostics", *Advances in Wound Care*, vol. 9, no. 5, pp. 245–263, 2020. DOI: 10.1089/wound.2019.0967.

[41] M. S. Nixon and A. S. Aguado, "8 - Region-based analysis", in *Feature Extraction and Image Processing for Computer Vision (Fourth Edition)*, M. S. Nixon and A. S. Aguado, Eds., Fourth Edition, Academic Press, 2020, pp. 399–432, ISBN: 978-0-12-814976-8. DOI: https://doi.org/10.1016/B978-0-12-814976-8.00008-7.

[42] M. Z. B. Jahangir, S. Akter, M. A. A. Nasim, K. D. Gupta, and R. George, *Deep learning for automated wound classification and segmentation*, Preprint, 2024.

[43] M. Alabdulhafith et al., "Automated wound care by employing a reliable UNet architecture combined with ResNet feature encoders for monitoring chronic wounds", *Frontiers in Medicine*, vol. 11, p. 1 310 137, 2024.

A Systematic Review of the Current Legal Position of eHealth Standards in Norway

Marianne Lodvir Hemsing
Department of Business, Strategy and Political Science
University of South-Eastern Norway
Kongsberg, Norway
e-mail: mlhemsing@gmail.com

Abstract— **This systematic review investigates how Norwegian courts engage with technical standards and European Conformity (CE) marking in legal disputes involving eHealth and medical software. Although European regulation increasingly relies on harmonised standards, a systematic screening of 36 legal decisions from the national case-law database Lovdata Pro (2015–2025) found only five cases referencing either standards or CE-marking, and in none were these references determinative. Standards appeared as supportive background at best, and CE-marks were invoked as compliance signals rather than legal authority. These findings suggest that, unless legally "activated" via regulation or contract, technical standards play little role in litigation. The study offers a legal baseline ahead of European Health Data Space (EHDS) rollout and provides recommendations for improving the enforceability of standards in Norway.**

Keywords-Health Technology; Medical Software; Standards; Policy in Digital Health.

I. INTRODUCTION

Digital-health software ranges from Electronic Health-Record (EHR) systems, diagnostic software, and mobile apps to Software as a Medical Device (SaMD) governed by EU (European Union) Regulation. eHealth solutions and digital health tools increasingly rely on standards for quality, safety, security, and interoperability - and regulatory alignment to function effectively within and across national health systems [1][2].

Three major EU instruments define the regulatory landscape for digital health. The Medical Device Regulation (MDR) [3], and the In Vitro Diagnostic Medical Devices (IVDR) Regulation [4], requires SaMD and diagnostic software to be marked according to European Conformity (CE), and conform to essential safety and performance requirements. The European Health Data Space (EHDS) Regulation [5], effective since March 2025, establishes interoperability obligations for EHR systems and health-data access services.

European Technical standards are drafted by consensus by the formal European Standardization Development Organizations (SDOs) European Committee for Standardization (CEN), European Committee for Electrotechnical Standardization (CENELEC), and European Telecommunications Standards Institute (ETSI). Standards remain voluntary unless incorporated in law, regulation or contract [6]. Once cited in the Official Journal of the European Union (OJEU), they become harmonised standards [7] and confer a rebuttable presumption of conformity to EU regulation [8]. There are as of September 2025, 44 harmonised standards related to MDR and IVDR (not all related to eHealth).

The standardization landscape is complex, as of June 2025 ISO's Technical Committee for Health Informatics had published 242 standards [9], and CEN's had published 118 [10]. In addition, there are standards from the other SDOs and standardization bodies outside the formal European standardization system, such as Health Level Seven (HL7). EU's New Legislative Framework [11] are based on legislators drafting "essential requirements," while the European Standardization bodies supply detailed solutions. The European Standardization Strategy [6] reinforces this model.

Several studies and rulings have highlighted the complex relationship between technical standards and legal transparency in the EU, and the blurred line between hard and soft law. In the *Public.Resource.Org* case [12], the General Court of the European Union ruled that harmonised standards incorporated into EU law must be publicly accessible, as they form part of the legal order. This decision underscored growing concerns about the accessibility of legal norms developed through private standardisation bodies.

Building on this theme, researchers have argued that the paywalled nature of harmonised standards poses a structural barrier to their legal enforceability and public legitimacy [13]. This suggests that unless such standards are freely available and embedded into binding legal texts, they are unlikely to feature prominently in litigation or regulatory practice.

In Norwegian law, national standards are referenced in the Regulation on IT Standards in Health and Care Services [14], which mandates the use of specific standards for interoperability, messaging, and security in eHealth systems. In addition, the Norwegian Product Control Act [15], which establishes a general duty of care for safe products and technologies, identifies adherence to national or EU harmonised standards as an indicator of responsible practice.

This creates a potential legal basis for invoking standards in negligence claims.

Menon Economics found in 2022 that references to standards are becoming more common in Norwegian regulations [16], yet there remains limited understanding of how courts treat such references in actual legal disputes.

Previous research suggests that standardization that relies on informal, consensus-driven public–private models, may hinder downstream legal enforceability in Norway [17]. Furthermore, research by Lindøe et al. [18] have identified three necessary legal hooks for standards to be influential in legal proceedings in Norway (the research did however not consider harmonized standard separately or eHealth specifically):

(1) explicit reference in contracts,

(2) incorporation into regulations or delegated law, and

(3) use in negligence assessments to define reasonable conduct.

No study has examined how Norwegian courts treat National and European standards in digital health disputes specifically. This paper provides a systematic legal review of Norwegian court practice on e-health standards, based on analysis of legal decisions from public legal records in Norway from the last decade. The aim is to determine whether, when, and how courts cite or rely on standards, and to offer a baseline for evaluating EHDS implementation in future litigation.

The remainder of this paper is structured as follows: Section 2 details the quantitative and qualitative method, Section 3 presents and discuss the findings, limitations and future research. Section 4 draws the conclusions.

II. METHODOLOGY

This study followed the PRISMA 2020 guidelines for systematic review [19]. The checklist, systematic review protocol, and strategy (including search terms, keyword logic, scripts, and classification scheme), is available at [20]. All legal data were sourced from Lovdata Pro, Norway's official case-law repository, which contains full-text judgments across all national court levels [21].

To capture the full scope of court decisions related to eHealth software, medical device software, and digital health technologies, two full-text Boolean searches were designed. These combined key words using logical AND/OR across two thematic fields (one term related to "software," one related to "health"). Searches were case-insensitive and used open-ended wildcards to capture inflectional forms. The time window was set from 1 January 2015 to 31 March 2025.

The search was structured as follows:

- Set 1 (software terms): technology*, informatics*, health record*, app, application*, software*
- Set 2 (health terms): health*, ehealth*, medicine*

The Boolean operator AND was applied between Set 1 and Set 2, requiring at least one keyword from each set to appear in the full text.

The dataset was in the initial analysis coded using a Python Script with a predefined coding scheme, as described in Table 1. Negatives were excluded from the dataset.

TABLE I. CASE CLASSIFICATION

Code	Description
eHealth	Case related to eHealth, Health Informatics, Medical app or technology
Standard	Cites an International, European or Norwegian Standard
MDR/IVDR/ CE	Cites the MDR, IVDR, references CE-Marking, references Harmonised Standards or OJEU

The following exclusion criteria was used:

- First Exclusion: Judgments unrelated to eHealth, medical technology, or software used in healthcare (including EHRs, apps, SaMD, CE-marked technology etc).
- Second Exclusion: Judgments unrelated to MDR, IVDR, CE-Marking, National or European Standards or OJEU
- Unit of analysis: Each full-text court decision.

After the initial quantitative screen, an interpretive, in-depth review of the full-text of each case was performed, to understand how each reference entered the legal reasoning and what weight the court gave it. For the identified qualitative sample, Case ID and date, Case summary, gateway (contract clause, regulation, negligence, or factual background), outcome role (determinative, supportive, descriptive), and cited standard(s) were recorded.

Figure 1 shows the PRISMA flow of records through the systematic review.

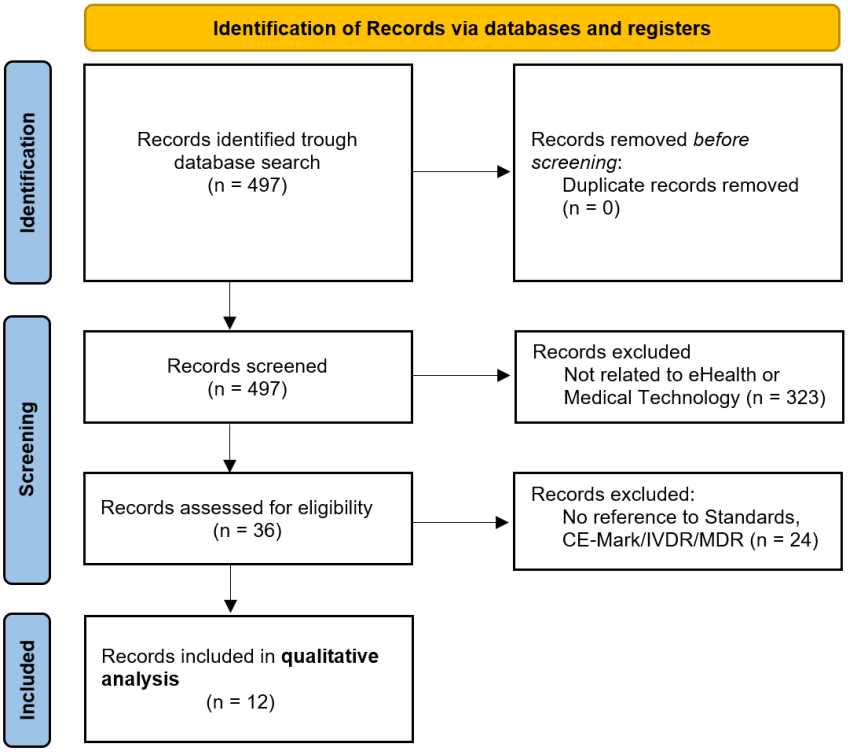


Figure 1. PRISMA Flow.

III. RESULTS

The database search yielded 497 judgments from the Norwegian legal database Lovdata Pro. After screening duplicates and irrelevant cases, a total of 36 legal decisions were identified as involving eHealth software, health information systems, or medical technologies. Among these, 12 cases were shortlisted as candidates for potentially involving references to technical standards or regulatory conformity frameworks.

A detailed qualitative analysis of these 12 cases revealed that only two decisions explicitly cited a formal technical standard. Notably, both of these cases also referenced CE-marking, suggesting that the standard citation was connected to product conformity documentation or procurement specifications. An additional three cases mentioned CE-marking without referencing any technical standard.

The remaining seven cases in the shortlist contained indirect references (such as mentions of regulatory requirements, compliance duties, or safety documentation) but did not directly cite or reference a standard or discuss CE-marking. The full list of the 12 cases is provided in Table 2, with cases referencing standards marked in Green and those citing CE-marking alone marked in Yellow.

These results suggest that while regulatory terminology may appear in health-tech litigation, direct engagement with standards remains rare.

TABLE II. CASE SUMMARY

Case no.	Subject	Standards in the reasoning
1	EPJ privacy dispute (2024)	Court reviews General Data Protection Regulation (GDPR) & Norwegian privacy rules, CE-Marking discussed
2	Proton Therapy System procurement (2023)	Addresses tender law; standards never cited. CE-Marking part of dispute
3	Online patient portal (2022)	Liability question decided on negligence; standards not referenced.
4	Wellness-app tax case (2022)	VAT classification only; standards not referenced.
5	Hospital IT-system outage (2023)	Focus on employer liability; standards not referenced.
6	Tele-medicine platform (2022)	Contract damages; standards not referenced.
7	Forensic phone-extraction tool used on medical professionals' communication (2022)	Proportionality & evidentiary law; standards not referenced.
8	Biofeedback fitness device (2021)	VAT issue; Standards and CE-Marking is cited
9	Bone-cement system (2021)	Product-liability; CE-Marking is cited
10	Anaesthesia equipment failure (2020)	Liability assessed via expert opinion; Standards and CE-Marking cited
11	Pacemaker follow-up system (2016)	Medical negligence claim; standards not referenced.
12	Medical equipment for local health care centre (2015)	Local contractual disputes; standards not referenced.

A. Two cases citing standards

The two cases citing standards are summarized in Table 3. Both cases mention internationally recognised medical-device standards and CE-marking, used as background evidence. The standards cited are in both cases EU harmonised standards, i.e., demonstrating conformity to EU Regulation (in this case MDR).

In neither judgment does conformity (or non-conformity) with those standards decide the legal result; the norms serve only as expert background or to show baseline regulatory approval.

Standards are informative, not determinative. The rulings rely on tax and negligence principles rather than on compliance or breach of the cited norms.

TABLE III. STANDARD CASES

Point	Case 8	Case 10
Type of dispute	VAT & civil-liability case about a bio-feedback / electro-stimulation fitness device ("Bailine method").	Product-liability / negligence case about an anaesthesia workstation that allegedly malfunctioned.
Standards invoked	Expert for the device owner cites ISO 13485 as proof the manufacturer operates an approved quality-management system.	Experts cite IEC 60601-1 (electrical safety) and its EMC collateral IEC 60601-1-2 to describe the minimum design-safety level for the workstation.
CE-marking	Mentioned once: the Bailine apparatus is CE-marked as a class-I device.	Mentioned twice: the workstation carried a CE-mark and Declaration of Conformity.
Weight given to the standards	Court's outcome (VAT classification) does not turn on ISO 13485; standard is noted but not analysed.	Court's finding (negligence) does not hinge on IEC 60601; standard is illustrative of good practice, not determinative.

B. Three cases citing CE-mark

Across the three cases, summarized in Table 4, CE-marking is raised only as baseline regulatory compliance. In all three cases it is cited to show that the product had a formal Declaration of Conformity for EU Regulations (MDR).

The courts treat the CE-mark as necessary but not sufficient. Each judgment acknowledges that CE-mark is a minimum legal threshold, yet it does not settle the central question (privacy breach, procurement legality, or product defect).

No judgment turns on a finding of CE non-compliance. In short, the three cases use CE-marking as background evidence of market approval, but the mark itself never drives the outcome, and no other technical standards are cited.

TABLE IV. CE-CASES

Case no.	Why CE-marking is mentioned	How the court treats it
1	Defendant hospital notes that the module is *"CE-marked as a Class IIa medical device."*	Court accepts that CE shows formal EU conformity but rules on GDPR/consent issues; CE is not part of the legal test.
2	Tender documents required every offered device to be CE-marked. Losing bidder claimed the winner lacked final CE paperwork.	Court finds the winner could submit missing certificates after award; CE is a procedural tender condition, not a ground to annul the contract.
9	Manufacturer stresses that the cement kit was CE-marked under MDD 93/42/EEC.	Court notes the mark but decides liability on causation/expert evidence; CE carries no decisive weight.

C. Discussion

1) The Peripheral Role of Standards in Norwegian Case Law

This systematic review found that only two of the 36 eHealth-related legal cases referenced a technical standard, and both of these also cited CE-marking. In neither instance did the standard serve as a decisive factor in the court's reasoning. Instead, courts resolved disputes based on general legal doctrines, such as negligence, contract interpretation, or procurement law.

This confirms the pattern observed by Lindøe et al. [18]: technical standards tend to shape legal reasoning only when they are linked to one of three legal gateways:

1. Explicit contract clauses (e.g., references in procurement tenders or service agreements),
2. Regulatory incorporation (e.g., CE-mark),
3. Negligence benchmarks (e.g., evidentiary use to define "reasonable care").

Absent these anchors, standards play at best a descriptive or supportive role. They may appear as evidence of good practice or industry norm, but not as legal authority in themselves.

2) CE-Marking: Binding, Visible, but Legally Passive

CE-marking appeared in five of the 36 analysed cases, more often than references to technical standards. In all five, CE-marking was acknowledged as proof of regulatory conformity. However, in no case did the court treat the CE-mark as determinative for liability, dismissal, or award of damages.

Courts appear to treat CE-marking as a higher-order legal norm than any individual standard. It is:

- Conferred by law, as required by the MDR and IVDR
- Presumed to indicate compliance with essential requirements, and
- Frequently cited in litigation as evidence of market access or eligibility.

Nonetheless, CE-marking remains procedurally visible but legally passive. In procurement cases, it functions as a

formal requirement. In product liability or privacy cases, it confirms baseline regulatory status, but courts still ground their decisions in traditional doctrines of causation, contractual breach, or data protection law.

As Volpato [8] notes, harmonised standards confer only a presumption of conformity unless incorporated into law. CE-marking may be invoked in litigation, but it rarely shifts outcomes without additional legal support.

3) *Why Courts Rarely Engage Directly with Standards*

While the Menon Economics report [16], documents an increase in regulatory references to standards across Norwegian legislation, the findings of this review suggest that such references rarely translate into legal reasoning or judicial outcomes unless standards are explicitly invoked through regulation, contract, or negligence frameworks.

The absence of standards in most decisions may reflect structural and procedural features of judicial reasoning:

- Deference to higher-order sources: Courts prioritize statutes, contracts, and regulatory instruments over third-party norms like standards.
- Lack of formal legal status: Most standards are non-binding unless cited in law or incorporated by contract. As Heyerdahl [17] shows, even nationally supported standardization efforts in Norway may operate outside formal legal channels, limiting their ability to shape judicial reasoning.
- Access barriers: Many technical standards are paywalled, hindering their citation and judicial consultation.
- Technical complexity: Standards often require domain-specific interpretation. Judges may prefer expert testimony or official guidance instead.

Together, these factors may explain some of the reasons why courts, even in technically regulated sectors, engage only superficially with formal standards unless they are "activated" through legal incorporation.

4) *Implications for Regulators, Litigators, and Industry*

The findings highlight a broader policy challenge; if courts do not engage with standards directly, even when invoked in digital health, the expected legal alignment under the EU Regulation may fail to materialize unless legal instruments and contracts explicitly operationalise them.

This limited judicial engagement with standards and CE-marking has important consequences:

- For regulators: Forthcoming EHDS common specifications must be embedded in binding instruments (e.g., delegated acts, procurement law) if they are to influence future litigation.
- For procurers and vendors: To give technical norms contractual force, actors should cite specific technical standards) in tenders and contracts.
- For litigators: CE-marking should not be assumed sufficient to establish compliance. Where relevant, standards should be referenced directly in pleadings and supported by expert interpretation.
- For standardisation bodies: The *Public.Resource.Org* ruling by the EU General Court calls for greater transparency in public access to harmonised standards, to enable legal analysis and citation.

Without these steps, courts will continue to default to general doctrines, and technical standards, however sophisticated, may remain silent in legal practice.

IV. CONCLUSION AND FUTURE WORK

A. *Limitations and Future Work*

The dataset from the Norwegian database Lovdata Pro excludes unpublished settlements and administrative market-surveillance measures; it therefore captures only disputes that reached the courtroom. Because all cases appeared in distinct factual settings, caution is needed before generalising.

Screening may have missed records that employed atypical terminology; Keyword-based scraping may miss cases that describe software obliquely. However, the highly specific keyword-set and automated full-text search mitigate this risk.

Further research should replicate this study in some years in the future, after the EHDS Regulation and its first harmonised standards are in force, to measure any uptake shift. A survey for litigants and judges on whether pay-walled standards deter citation in pleadings, could potentially test the transparency hypothesis.

B. *Conclusion*

This systematic review shows that Norwegian courts seldom cite technical standards or CE-marking when adjudicating disputes in digital health. In a decade's worth of cases, only five referenced either, and none treated these references as legally determinative. Instead, courts relied on established doctrines of contract, negligence, or procurement law, treating standards and CE-marks as background context rather than binding authority.

These findings underscore the limited traction of technical standards in Norwegian legal reasoning, despite their central role in the MDR, IVDR, and broader EU digital health strategy. Unless standards are explicitly incorporated into law, regulation, or contract, they are unlikely to play a decisive role in courtroom outcomes.

For policymakers and industry, this highlights the need to anchor technical norms, such as those soon to emerge under the EHDS, in binding instruments and contractual frameworks if they are to have legal bite. CE-marking, while more frequently cited than individual standards, is also treated as a procedural or evidentiary formality, rather than a substantive safeguard or evidence in litigation.

This reinforces two strategic insights:

- To give technical norms legal bite, anchor them in delegated regulations or procurement frameworks.
- To ensure contract enforcement, cite specific standards directly in agreements.

Absent such hooks, standards may remain invisible in litigation.

The study provides an empirical baseline for the legal treatment of eHealth standards in Norway prior of the EHDS implementation. Future research should revisit this landscape as new EU requirements take effect, and explore how accessibility, transparency, and legal embedding of standards may shift the role of technical norms in judicial decision-making.

REFERENCES

[1] Z. Wong, Y. Gong, and S. Ushiro, "A pathway from fragmentation to interoperability through standards-based enterprise architecture to enhance patient safety," *npj Digital Medicine,* vol. 8, no. 1, p. 41, January 2025.

[2] J. Scheibner, M. Ienca, J. Sleigh, and E. Vayena, "Benefits, Challenges and Contributors to Success for National eHealth Systems Implementation: A Scoping Review," *Journal of the American Medical Informatics Association : JAMIA,* no. 28(9), p. 2039–2049, 2021.

[3] European Union, "Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation(EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC", January 2025.

[4] European Union, "Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU", January 2025.

[5] European Union, "Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847," March 2025.

[6] European Commission, "COM/2022/31 An EU Strategy on Standardisation - Setting global standards in support of a resilient, green and digital EU single market," February 2022.

[7] European. Commission, "Harmonized Standards," 2025. [Online]. Available: https://single-market-economy.ec.europa.eu/single-market/european-standards/harmonised-standards_en., retreived: September, 2025.

[8] A. Volpato, "The legal effects of harmonised standards in EU law: From hard to soft law, and back?" i *The Legal Effects of EU Soft Law*, Edward Elgar, 2023, pp. 193-212.

[9] ISO, "ISO/TC 215 International Technical Committee Health Informatics," 2025. [Online]. Available: https://www.iso.org/, retreived: September, 2025.

[10] CEN, "CEN/TC 251 European Technical Committee for Health Informatics," [Online]. Available: https://www.cencenelec.eu/, retreived: September, 2025.

[11] European Union, "Regulation (EC) No 765/2008 of the European Parliament and of the Council of 9 July 2008 setting out the requirements for accreditation and market surveillance relating to the marketing of products and repealing Regulation (EEC) No 339/93", July 2021.

[12] European Union, "Case T-185/19 Public.Resource.Org, Inc. & Right to Know CLG v European Commission, Judgment of the General Court (Fifth Chamber, Extended Composition)" July 2021.

[13] E. Tzoulia, "Harmonized Standards in the Public Domain? Better Not …," *IIC - International Review of Intellectual Property and Competition Law,* vol. 56, nr. 4, p. 692–712, 2025.

[14] Stiftelsen Lovdata, "FOR-2015-07-01-853 Norwegian Regulation on IT Standards in Health and Care Services (Forskrift om IT-standarder i helse- og omsorgstjenesten)", 2015.

[15] Stiftelsen Lovdata, "LOV-1976-06-11-79 Norwegian Product Control Act (Produktkontrolloven)", 2021.

[16] Ø. Vennerød et al., "Reference to Standards in Norwegian Regulations (Henvisning til Standarder i Norsk Regelverk)", *Menon Economics,* 2022.

[17] A. Heyerdahl, "Standardising policy in a nonstandard way: a public/private standardisation process in Norway" *Journal of Public Policy,* nr. 4, p. 761–790, 2023.

[18] P. Lindøe, J. Kringen, and G. S. Braut, «Regulation and standardization: perspectives and practice (Regulering og standardisering : perspektiver og praksis)», *Universitetsforlaget*, 2018.

[19] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews,» *BMJ,* vol. 372, p. n71, 2021.

[20] "Figshare Dataset," [Online]. Available: 10.6084/m9.figshare.29610470.

[21] Stiftelsen Lovdata, "Lovdata Pro," [Online]. Available: https://lovdata.no/pro/.

Using a Large Data Model Explorer to Maintain a Healthcare Information System

Rubén Martínez Martínez*, Francisco Javier Bermúdez Ruiz*
Manuel Campos Martínez*†, José Manuel Juarez Herrero*†
*MedAI Lab, University of Murcia, Spain
†Murcian Bio-Health Institute (IMIB-Arrixaca)
Email: {ruben.martinez11, fjavier, manuelcampos, jmjuarez}@um.es

***Abstract*—Maintaining large scale healthcare information systems poses significant challenges due to their evolving complexity and sparse documentation. This paper presents a methodology and supporting tool for exploring and understanding extensive data models, motivated by and applied on the Wise Antimicrobial Stewardship Support System (WASPSS) application, an antimicrobial stewardship support system in production across several Spanish hospitals. The proposed tool enables incremental, interactive visualization of the Java Persistence API (JPA) based domain model in WASPSS, including inheritance and relationships, facilitating maintenance and onboarding. Key features include partial view saving, entity tagging, and dynamic inheritance propagation. An evaluation based on the Technology Acceptance Model (TAM) confirms the tool's perceived usefulness and ease of use among software professionals, supporting its applicability in real world maintenance workflows and its potential reuse in other healthcare contexts.**

Keywords-WASPSS; healthcare system; maintenance; large data model; model explorer.

I. Introduction

Hospital acquired infections, also known as Nosocomial Infections (NI), have become a serious public health problem. They are defined as infections that appear 48 hours after hospital admission and are often caused by multidrug resistant bacteria, which have lost susceptibility to common treatments, increasing their spread, severity, and lethality [1] [2]. Misuse of antibiotics contributes to the emergence of new resistances, limiting available therapeutic options [3]. NIs increase morbidity, mortality, and hospital costs due to more complex treatments and prolonged stays [4].

In the context described above, the WASPSS tool [5] (*Wise Antimicrobial Stewardship Program Support System*) emerges, as a decision support platform for Antimicrobial Stewardship Programs (ASP). WASPSS was initially developed by the University of Murcia and the Getafe University Hospital, where it was implemented in 2015, and has since been used daily by ASP teams as part of the National Plan against Antibiotic Resistance. In 2018, it began piloting in seven hospitals across various autonomous communities. Currently, WASPSS is in production in all hospitals of the Basque Country and the Murcia Health Service of the Region of Murcia.

WASPSS is a web application developed following the Model-View-Controller pattern and based on standard Java technologies. It offers interactive interfaces tailored to different ASP team profiles. The business logic is organized through façade objects, encapsulating specific functionalities and accessing persisted data via data access objects. Persistence is managed by Java Persitence API (JPA) and a PostgreSQL relational database, which serves as the central information core for all modules. WASPSS integrates with hospital systems via an HL7 interface, a widely used standard in clinical environments for real time information exchange. Finally, the knowledge modules use Drools, allowing clinical rules to be applied to hospital data to generate alerts and support clinical decision making by the ASP team.

The WASPSS project has been under development for over 12 years. Over this time, its codebase has grown organically, incorporating new features and adapting to changing business requirements. However, the project documentation has not scaled at the same pace as the application. This hinders onboarding new developers and maintaining the system. This growth has far outstripped the capacity of the existing static documentation, which has not been systematically updated at the code's pace. The commitment of the developing team to software quality and ongoing development makes the effort to enhance the application maintainability, both meaningful and motivating. This will, in turn, facilitate to add new features, fix bugs, or make improvements without significant prior knowledge of the code. In addition, this also will speed up development, reduce the risk of errors, and ease knowledge transfer between developers. Examining further, the project includes over 1,000 Java classes, many organized in inheritance hierarchies and complex relationship networks. Additionally, the JPA domain model includes over 200 Java classes, and the database contains 166 tables across 9 schemas. These figures contrast sharply with the sparse documentation available. A good percentage of the code is undocumented, and there are few tool reports. Maintenance tasks thus become difficult due to the challenge of understanding the underlying model. Furthermore, existing market tools for exploring information system data models face difficulties in generating complete visual reports or diagrams containing all the model information. This feature is unsuitable for very large data models, as in the WASPSS project. In other words, trying to visually represent hundreds of entities in a single diagram becomes unintelligible to humans, with graphics that overlap and hide much of the information. All of this highlights the challenge of maintaining the tool, both for reasoning about and understanding the underlying domain model that drives the system and for grasping the structure of resources comprising the WASPSS project itself.

Note that in the development of an information system, the domain model represents the core of the software since it encapsulates the fundamental knowledge, rules, and processes of the business. Together with requirement specifications, this

 ISBN: 978-1-68558-312-5

model forms the foundation upon which the system is built and evolves. In approaches like Domain Driven Design [6], it is recognized that the true complexity of software lies not in the technology but in the domain itself, reflecting the design of the system. Thus, the domain model is not just another system component, but rather the central structure guiding its design. This work proposes a methodology for discovering and exploring a large data model in sparsely documented large scale information systems. The methodology is supported by a tool that allows loading the model information to reason about it incrementally and progressively, discovering the domain entities structuring the application and enabling development teams to work with customized views that allow them to segment the part of the model they want to work on. Since the exploration methodology relies on a visualization tool that progressively reveals the model, it is important that the information displayed matches the graphical context at each stage. This implies the need to visually propagate inherited information between entities in an object oriented paradigm–based model. The information propagated through inheritance will be displayed in some entities or others, depending on which entities are currently visualized (and the inheritance involved). Specifically, it means that our system parses each JPA entity, including its attributes, inheritance hierarchy, and relationships, to produce a normalized representation. Each entity is then rendered as a node in the interactive graph, while attributes and relationships are displayed as labeled fields and directed edges, respectively. Inheritance is represented visually by propagating the properties of parent classes to their child nodes when the parent is not explicitly shown. This mapping ensures that developers can incrementally reveal and reason about the full domain model without overwhelming diagrams. The current version of the tool offers only a basic level of customization for visualization. For instance, users cannot yet selectively choose which specific relationships to display or hide, nor can they freely reposition individual relationship edges within the diagram.

To guide our work, we formulate the following research questions: (i) how can the exploration and comprehension of a large, sparsely documented data model in a healthcare information system be improved?; (ii) what functionalities are required in a model exploration tool to support effective maintenance and onboarding of developers?; and (iii) to what extent is the proposed tool perceived as useful and easy to use by software professionals?. Research questions are aligned to the main contributions of this work: the proposal of a methodology for exploring large-scale, poorly documented data models; the development of a supporting tool; and the validation of this tool by end users.

In Section II, we enumerate the main tools that address similar problems. Section III presents the outcomes obtained from the proposed approach. Next, in Section IV, we evaluate how useful the tool is perceived. Finally, in Section V, we summarize the key findings and outline potential directions for further research.

II. Related work

Previous research has also explored interactive visualization techniques to manage the complexity of large healthcare information systems. [7] proposed *Owlready*, an ontology-oriented programming framework for biomedical ontologies that enables incremental exploration and automatic classification of complex domain models, addressing similar challenges of scalability and maintainability. [8] introduced *Health Timeline*, a timeline-based visualization that allows clinicians to progressively explore patient records, demonstrating that focused, interactive views improve comprehension of large clinical datasets. Likewise, [9] presented a richly interactive exploratory visualization tool for electronic health records, highlighting the value of dynamic filtering and navigation for handling extensive medical data. These works support the need for incremental, user-centered visualization approaches, which our proposed JPA model visualizer extends to the maintenance of large-scale healthcare data models.

In Java environments, complex data models with JPA entities, relationships, and inheritance are difficult to grasp directly from code as their size increases. To address this, several tools and plugins provide graphical visualization of domain models from JPA entities or database tables. Next, a brief review of notable tools is enumerated: (i) **JPA Diagram Editor (Eclipse Dali)** [10]: Free Eclipse plugin for visual JPA editing, but limited to IDE use and lacks dynamic, interactive diagrams; (ii) **Hibernate Tools / Mapping Diagram** [11]: Shows entity mappings and relationships in read-only form; offers basic hiding/collapsing of connections; (iii) **Jeddict (JPA Modeler)** [12]: NetBeans plugin for creating and editing entities with code–diagram sync and reverse engineering from databases; (iv) **IntelliJ IDEA Ultimate** [13]: Provides a persistence view for JPA entities with simple navigation and layout, but limited editing options; (v) **JPA Buddy** [14]: IntelliJ plugin focused on code generation with minimal visualization features; (vi) **Generic modeling tools** (DBeaver, SchemaSpy, etc. [15], [16]): Can draw database/UML diagrams but do not handle JPA annotations or source-level models.

The conclusion after analyzing various tools is that, although each one offers certain features that may be of interest, none of them allow for incremental and interactive exploration starting from an initial entity or for automatic management of inherited attribute propagation features we consider essential for understanding a large scale model. Moreover, most of them do not allow saving persistent partial views of the model, nor do they offer a smooth integration between visual editing and an interactive, dynamic diagram.

III. Results

The solution proposed in this work is based on offering, through a support tool, the ability for developers to progressively and interactively explore the application data model. Additionally, the information displayed in the visualization tool must be dynamic and adapted to the entities shown on screen, according to the data propagation behavior inherent to inheritance relationships in an object oriented data model.

Inheritance is a mechanism that allows a class (called a subclass or child class) to inherit the properties (attributes) and methods (behaviors) of another class (called a superclass or parent class). In this context, when entities (classes) are related through properties (for example, a property referencing another class), those relationships are also propagated to the subclasses. That is, if a superclass has a property managing a relationship with another entity (e.g., a list of related objects), the subclass will also inherit that relationship, allowing for the management of relationships between objects through inherited properties. Therefore, we propose the design and implementation of a graphical visualization tool that enables any member of the development team to explore and navigate the JPA model interactively, displaying both the own class and inherited information depending on the selected visualization elements. A diagram based on UML class diagrams [17] will be used, with slight visual modifications to the standard UML notation. Since we are only interested in the structural information of the data model, the current version of the tool omits information about methods (functions) of each entity in favor of a clearer visualization. Figure 1 shows an example of entities with attributes and relationships.

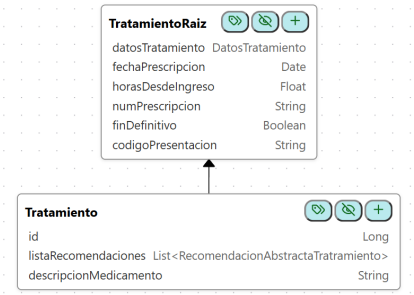


Figure 1: Example of entities with attributes and one inheritance relationship.

A. Main functionalities

The basic functionalities of the application must be as follows:

- Ability to progressively and interactively explore the model to discover relationships and inheritance from the entities themselves.
- Ability to dynamically add and remove entities from the visualized diagram. Each time the diagram is modified, our visualizer will show inherited information consistent with the current state of the diagram.
- Ability to save and retrieve partial views of the model. This allows designing visualizations focused on specific parts of the model, facilitating understanding and reasoning.
- Ability to categorize entities using tags. This allows filtering the model based on the tags applied to entities, enabling the display of all entities with a specific set of tags selected by the developer with a single click.

Although the visualizer developed in this work was designed with the primary goal of integration into the `WASPSS` project ecosystem, it is important to highlight that its implementation is not tightly coupled to that system. The modular architecture of the visualizer allows it to be reused in other contexts, as long as the data model can be adapted to be consumed by the visualizer. The only requirement is that the model be object oriented, i.e., composed of entities with attributes, relationships, and hierarchical structures similar to those of an object oriented system. This domain independence opens the door for the visualizer to be applied in other medical information systems beyond `WASPSS`, reinforcing its value as a generic tool for exploring complex data models.

B. Architecture of the Visualization Tool

The architecture consists of two functional units (see Figure 2):

- A parser responsible for analyzing and extracting information from the data model based on project resources (source code).
- An interactive visualizer that displays the information extracted from the data model.

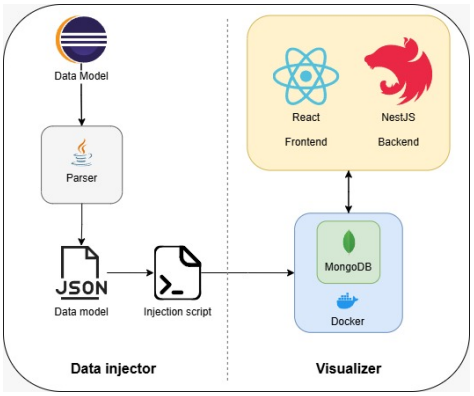


Figure 2: Architecture of the Visualization Tool.

C. Extraction and Normalization of the JPA Model

The solution arises from the need to visually represent a complex data model. To this end, a parser has been developed that automatically analyzes the project's source code and extracts structured information about JPA entities and their relationships. This process identifies each entity, its

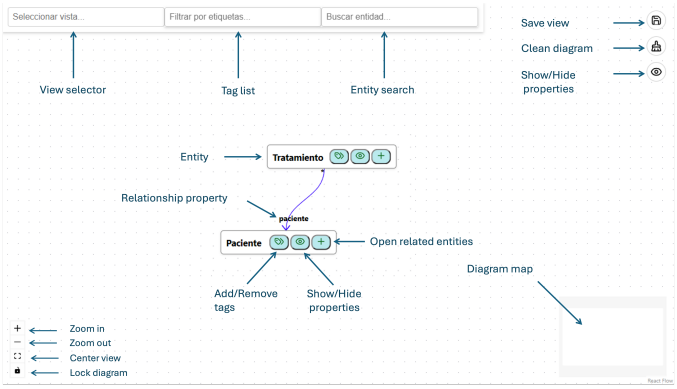


Figure 3: User Interface of the Visualization Tool.

inheritance hierarchy, and the different types of relationships. The output is a set of structured representations (one per entity) stored in a avaScript Object Notation (JSON) file, which can be processed by subsequent components of the solution. Thanks to the use of JSON as an intermediate format, the visualizer is not directly coupled to the `WASPSS` project and can be reused in any other information system for which source code is available. This means that, as long as a properly structured JSON file is available, the tool can be reused without modifications. This promotes reusability of the visualizer in other contexts and ensures greater flexibility and scalability. The JSON metamodel defines the basic properties for an entity identified by the parser. These are: `className`, `parentClassName` (the class from which it inherits, if any), `packageName` (the full class path), and a collection of `fields`, which are defined by `fieldName`, `fieldType`, `isRelationship` (indicating whether the attribute is a simple attribute or whether it represents a relationship with another entity), `relationshipType` (indicating the cardinality of the relationship) and `targetEntity` (indicating the entity with which it is related).

D. Storing Information

The parser analyzes the project source code (i.e. the JPA classes) to extract the model representation in JSON format. This JSON file is then loaded via a script into a MongoDB database (running in a Docker container). The choice of database is due to its ease of working directly with JSON documents. The visualizer uses the database to read the data model and persist interactive visualization information from developers, managing views and tags.

Once the model is extracted, the data is inserted into a database via a script. This database, in addition to containing the complete list of model entities and their relationships, provides the visualizer with the capability to store user created views and entity assigned tags. Therefore, this component acts as the system persistence core and ensures data consistency throughout user interaction with the visualizer.

E. Interactive Model Visualization

The core functionality of the solution is a graphical interface that allows users to interactively explore the entity model. Through this interface, the user can:

- Visualize JPA entities as nodes in a dynamic diagram.
- Explore relationships between entities, visually represented as directed edges.
- Save partial views of the model for later retrieval.
- Tag entities to aid organization and filtering.

During usage, the visualizer automatically manages the visual graph's consistency, including the incorporation of inherited relationships when ancestor entities are not present in the current view.

F. Interface Elements

The entity visualizer interface includes various interactive elements designed to facilitate data model exploration and customization. At the top of the interface are three key components (see Figure 3):

- `View Selector`: allows the user to switch between different views previously created. A view is a set of saved entities that can later be restored.
- `Tag List`: provides tag based filtering to reduce the number of visible entities, showing only those associated with selected categories. This is useful for locating all entities belonging to a specific concept without searching one by one.
- `Entity Search`: offers a text field to search for a specific entity by name or navigate through the full list.

In the central workspace, entities are represented as nodes. Each node appears as a rectangle with three buttons, each providing specific functionality (explained below). Relationships between entities are shown as directed edges, labeled with the name of the relation (e.g., patient) and its cardinality. At the top right, three additional buttons provide global functionality:

- `Save View`: stores the current diagram layout as a new view, including entities, relationships, and their positions.
- `Clear Diagram`: removes all visible nodes and any selected views or tags from the current diagram.
- `Show/Hide Properties`: toggles the visibility of all entity fields globally.

A minimap is displayed in the bottom-right corner, offering a quick overview of the canvas and enabling faster navigation. Finally, in the bottom-left, there are navigation controls enabling: `Zoom in / Zoom out`, `Center view` and `Lock diagram` (temporarily disables interaction to prevent accidental changes).

Figure 4: Node example.

Each node consists visually of a header with the entity name and three buttons. As seen in Figure 4, from left to right:

- `Tag Button`: opens a dropdown for assigning or removing tags. Users can search, activate, deactivate, or create new tags. Any change triggers an call to sync the backend.
- `Show/Hide Fields Button`: toggles the visibility of the entity's attribute list via an internal boolean. It also respects the global "show all fields" toggle.
- `Connection Button`: arguably the most important one. It displays a dropdown of related entities not yet shown in the diagram. These relationships include explicit ones, inherited children, and non-visible ancestors. When a user selects an entity, handleSelectedEntity is called to add the node and update edges automatically, enabling interactive discovery of the data model.

G. User Experience Enhancements

Several technical improvements were added to enhance user experience:

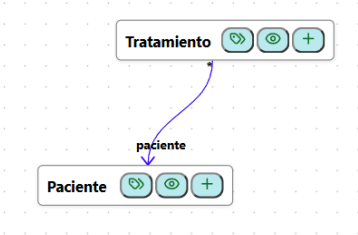


Figure 5: Simple relationship.

Figure 6: Multiple relationship.

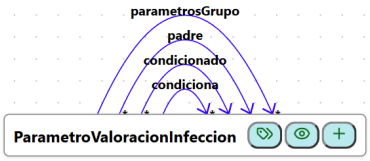


Figure 7: Reflexive relationship.

- When opening the tag creation modal, the focus is automatically placed in the input field.
- Scrolling inside the relationship dropdown does not affect canvas zoom.
- Clicking outside dropdowns closes them automatically (e.g., when navigating or using top filters).

H. Relationships

Figures 5, 6, and 7 illustrate simple, multiple, and reflexive relationships, respectively. Custom algorithms calculate the exact label and cardinality positions, accounting for node geometry and displacement. These implementations ensure uniform, clear rendering of semantic model elements. Three key elements are rendered per relationship:

- `Relation Label`: shown near the center of the edge; represents the field name of the association.
- `Source Cardinality`: marked with * for "ManyTo..." cases, shown near the source node.
- `Target Cardinality`: marked with * for "...ToMany" cases, shown near the target node.

All labels are interactive: clicking on any of them highlights the corresponding edge and its metadata (e.g., increasing size or changing color).

I. Saving and Managing Views

One key feature is the ability to save partial views of the data model. This allows users to focus on a subset of entities relevant to a specific task, hiding unrelated elements. Internally, each view stores exact canvas positions so when reloaded, the layout is reconstructed identically, maintaining the user's custom organization.

An interactive dropdown enables quick access to saved views:

- Typing in the search field dynamically filters available views.
- Selecting a view automatically loads its entities and their positions.
- Each view includes a delete button for direct removal.

J. Filtering by Tags

The main goal of tag based filtering is to allow users to instantly retrieve all entities marked with a certain tag. These tags can be user defined at any point. The tag dropdown appears when clicking the label icon in the nodes. It allows assigning existing or creating new tags. The tag creation modal includes validations to prevent empty or duplicate names.

IV. Evaluation

A validation of the tool was carried out by different users following the Technology Acceptance Model (TAM) [18], which allows us to understand how useful the tool is perceived to be, according to its purpose and the methodology defined in this work. We will describe the methodology used to evaluate the visualizer, as well as the results obtained and the conclusions drawn from them.

A. Technology Acceptance Model

TAM is one of the most established frameworks for analyzing how users adopt new technologies. This framework identifies two main determinants: perceived usefulness (the degree to which using the tool improves task performance) and perceived ease of use (the degree to which using the tool requires minimal effort) [18]. Therefore, this evaluation determines whether the tool meets two key hypotheses derived from TAM:

- H1. The application is simple and intuitive to use.
- H2. The application is perceived as useful.

Additionally, the model considers the user attitude toward the technology and their future intention to use it [18].

B. Evaluation Instrumentation

To conduct the evaluation of our visualizer, we designed an exercise [19] to be carried out using the application. Once the exercise was completed, the participants were asked to fill out a questionnaire divided into two sections:

- Demographics and experience: Age, gender, level of experience using computers, and experience with web applications.
- TAM Dimensions: Questions [20] grouped by perceived ease of use (items B1 to B9), perceived usefulness (items B10 to B16), attitude toward use (items B17 and B18) and intention to use (items B19 and B20) the visualizer.

C. Participant Demographics

The questionnaire was administered to 13 faculty members and staff from the Faculty of Computer Science at the University of Murcia. 61.5% of the participants were male and 38.5% female. The average age was 39.23 years. Most participants reported a high level of experience with both computer use and web applications.

D. Validation Results

Based on the responses collected, the mean values and standard deviation were calculated for each item using a 5-point Likert scale (1 = Strongly disagree, 5 = Strongly agree). The results obtained reflect an overall positive assessment of our visualizer by the participants. We can observe this in Table I.

TABLE I: MEANS AND STANDARD DEVIATIONS

Questions	Mean	Stand. Dev.	Questions	Mean	Stand. Dev.
B1	4.69	0.48	B11	4.92	0.28
B2	4.46	0.52	B12	4.46	0.88
B3	4.92	0.28	B13	4.85	0.55
B4	4.85	0.38	B14	4.23	1.09
B5	4.62	0.51	B15	4.46	0.66
B6	4.69	0.63	B16	4.92	0.28
B7	4.85	0.38	B17	4.85	0.38
B8	4.92	0.28	B18	4.69	0.63
B9	4.77	0.60	B19	4.69	0.48
B10	4.77	0.44	B20	4.77	0.44

For Perceived Ease of Use (items B1 to B9), a mean score of 4.75 was obtained, indicating that users found the visualizer easy and intuitive to use. The questions related to Perceived Usefulness (items B10 to B16) achieved a mean score of 4.66, highlighting that participants view the visualizer as a functional and valuable solution. The questions related to Attitude Toward Use (items B17 and B18) had a mean score of 4.77, indicating that participants believe using the visualizer is a good idea. Finally, the Intention to Use (items B19 and B20) received a mean score of 4.73, suggesting a strong willingness to use the visualizer in a real world context. The four values are close to 5, which corresponds to the rating *Strongly Agree*. These results allow us to conclude that our visualizer shows high levels of user acceptance.

V. Conclusion and Future Work

This work presents a methodology and tool to support the exploration and maintenance of large scale, poorly documented data models in healthcare information systems. Applied to the WASPSS platform, the tool enables incremental, inheritance visualization of complex JPA based domain models, significantly improving comprehension and maintainability. Our solution allows developers to interactively explore entities, manage customized views, and filter components using tags. The modular architecture ensures reusability beyond WASPSS, offering potential benefits for other object oriented medical systems. The positive results from the user evaluation, based on the TAM demonstrate that the tool is both intuitive and effective, confirming its value in real world development.

Future work will focus on extending the tool with collaborative features, model editing capabilities (such as the ability to show/hide relationships or reposition edge labels), and integration with automated documentation pipelines. In addition, we plan to extend the evaluation by increasing the number of participants and including professionals outside the academic environment, such as developers and IT staff from companies, in order to obtain a broader and more representative assessment of the tool.

Acknowledgment

This work was partially funded by the CONFAINCE project (Ref: PID2021-122194OB-I00) by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe", by the "European Union".

References

[1] A. Iqbal *et al.*, "Nosocomial vs healthcare associated vs community acquired spontaneous bacterial peritonitis: Network meta-analysis", *The American Journal of the Medical Sciences*, vol. 366, no. 4, pp. 305–313, 2023.

[2] R. B. McFee and G. G. Abdelsayed, "Clostridium difficile", *Disease-a-Month*, vol. 55, no. 7, pp. 439–470, 2009, Clostridium difficile: Emerging Public Health Threat and Other Nosocomial or Hospital Acquired Infections.

[3] WHO Regional Office for Europe and European Centre for Disease Prevention and Control, *Antimicrobial resistance surveillance in europe 2022 – 2020 data*, Copenhagen, 2022.

[4] R. E. Nelson *et al.*, "National estimates of healthcare costs associated with multidrug-resistant bacterial infections among hospitalized patients in the united states", *Clinical Infectious Diseases*, vol. 72, no. Supplement$_1$, S17–S26, Jan. 2021.

[5] B. Cánovas Segura, A. Morales, J. M. Juarez, M. Campos, and F. Palacios, "Waspss: A clinical decision support system for antimicrobial stewardship", in *Recent Advances in Digital System Diagnosis and Management of Healthcare*, K. Sartipi and T. Edoh, Eds., IntechOpen, 2020.

[6] E. Evans, *Domain-driven design: tackling complexity in the heart of software*. Addison-Wesley Professional, 2004.

[7] J.-B. Lamy, "Owlready: Ontology-oriented programming in python with automatic classification and high level constructs for biomedical ontologies", *Artificial Intelligence in Medicine*, vol. 80, pp. 11–28, 2017, ISSN: 0933-3657.

[8] A. Ledesma *et al.*, "Health timeline: An insight-based study of a timeline visualization of clinical data", *BMC Medical Informatics and Decision Making*, vol. 19, Aug. 2019.

[9] C.-W. Huang *et al.*, "A richly interactive exploratory data analysis and visualization tool using electronic medical records", *BMC medical informatics and decision making*, vol. 15, p. 92, Nov. 2015.

[10] Eclipse-Foundation, *Dali java persistence tools*, https://projects.eclipse.org/projects/webtools.dali, Last accessed: July 13, 2025.

[11] Red-Hat, *Hibernate tools*, https://tools.jboss.org/features/hibernate.html, Last accessed: July 13, 2025.

[12] Jeddict-Project, *Jeddict - jpa modeler and more*, https://jeddict.github.io, Last accessed: July 13, 2025.

[13] JetBrains, *Jpa in intellij idea*, https://www.jetbrains.com/help/idea/jpa-buddy.html, Last accessed: July 13, 2025.

[14] J. B. Team, *Jpa buddy — productivity tool for jpa/hibernate developers*, https://www.jpa-buddy.com, Last accessed: July 13, 2025.

[15] D. Corp., *Dbeaver — universal database tool*, https://dbeaver.io, Last accessed: July 13, 2025.

[16] S. Team, *Schemaspy — database documentation tool*, https://schemaspy.org, Last accessed: July 13, 2025.

[17] I. Jacobson, G. Booch, and J. Rumbaugh, *UML: The Unified Development Process*. Addison-Wesley, 2000.

[18] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology", *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989.

[19] https://github.com/fjavier-umu/healthinfo25/blob/main/exercise, Last accessed: July 13, 2025.

[20] https://github.com/fjavier-umu/healthinfo25/blob/main/questionnaire, Last accessed: July 13, 2025.

From Abstracts to Full Texts: The Impact of Context Positioning in LLM-Based Screening Automation

Elias Sandner
CERN
Geneva, Switzerland
e-mail: elias.sandner@cern.ch

Marko Zeba
University of Technology
Graz, Austria
e-mail: m.zeba@student.tugraz.at

Igor Jakovljevic
CERN
Geneva, Switzerland
e-mail: igor.jakovljevic@cern.ch

Alice Simniceanu
WHO
Geneva, Switzerland
e-mail: simniceanua@who.int

Luca Fontana
WHO
Geneva, Switzerland
e-mail: fontanal@who.int

Andre Henriques
CERN
Geneva, Switzerland
e-mail: andre.henriques@cern.ch

Andreas Wagner
CERN
Geneva, Switzerland
e-mail: andreas.wagner@cern.ch

Christian Gütl
University of Technology
Graz, Austria
e-mail: c.guetl@tugraz.at

***Abstract*—Screening for relevant research is among the most time-intensive phases of a Systematic Review (SR), significantly impacting its timeliness and resource requirements. Automation through Large Language Model (LLM) promises substantial efficiency gains, potentially reducing the human screening workload and mitigating the risk of reviews becoming outdated prior to publication. Existing research has primarily explored LLM applications in Title & Abstract (TiAb) screening, achieving promising sensitivity but limited investigation into Full-Text (FT) screening. This study extends the 5-tier prompting approach, originally developed for TiAb screening, to FT screening. An experimental evaluation was conducted using the LLaMA 3.1 8B model on five real-world SR datasets. Two FT prompting strategies were tested: one that directly adapted the 5-tier TiAb approach to FT screening, and another addressing the known 'lost-in-the-middle' phenomenon by positioning eligibility criteria before and after the full text. Findings indicate that providing FT context improves workload reduction considerably, nearly doubling it in some cases, though sensitivity slightly decreased compared to TiAb screening. Notably, positioning eligibility criteria both before and after FT significantly improved performance, highlighting the importance of the prompt structure. These results demonstrate that careful prompt engineering enhances LLM effectiveness in FT screening, balancing the critical trade-off between sensitivity and workload reduction. Overall, this research underscores the potential of LLM-based FT screening, providing valuable insights into prompt optimization for systematic review automation.**

Keywords-systematic review; screening automation; full-text screening; LLM.

I. INTRODUCTION

The screening process, where researchers evaluate the relevance of papers to a predefined research question based on eligibility criteria, is one of the most time-consuming aspects of a Systematic Review (SR) [1]. The automation of this process is essential for reducing human workload, thereby enabling timely, high-quality evidence-based research, particularly in time-sensitive situations or projects with limited resources. Furthermore, automation addresses the challenge that some SRs become outdated already by the time they are published [2].

Several studies have evaluated Large Language Models (LLMs) for screening automation in Title & Abstract (TiAb) and, more recently, Full-Text (FT) screening. However, only a few of the LLM-powered TiAb screening approaches are extensible for FT screening.

Since token limits no longer restrict LLM-based screening automation to the TiAb phase, the traditional separation between TiAb and FT screening can be reconsidered in the context of automated approaches. The 5-tier prompting approach, which acts as a prefiltration mechanism by removing records where the LLM is highly confident that the eligibility criteria are not met, has demonstrated promising results for TiAb screening [3]. Given its inherent scalability, this study focused on extending and evaluating the 5-tier prompting method for FT screening. The following research question is addressed through an experimental evaluation in which two FT prompting strategies are benchmarked against the original TiAb prompt, using LLAMA 3.1 on five real-world datasets:

Does providing the FT as additional context during screening yield higher sensitivity and greater workload reduction compared to LLM-powered TiAb screening?

The remainder of the paper is organized as follows: In Section II, the steps needed to conduct a SR are described, followed by a summary of related work in LLM-powered screening in SRs. Furthermore, Section III describes how the experiments have been conducted and which SRs were used for evaluation. The results are presented in Section IV, while Section V answers the research question by interpreting them. Finally, Section VI concludes this study and points out potential

 ISBN: 978-1-68558-312-5

future work based on the findings of this study.

II. Background and Related Work

SRs follow a structured approach by i) retrieving potentially relevant primary research, ii) evaluating the eligibility of those candidate studies, and iii) synthesizing the relevant findings [4].

In the first phase of a SR researchers define a research question. Based on this, the corresponding eligibility criteria, which are divided into inclusion and exclusion criteria, are defined. An insensitive search string is used to retrieve potentially relevant papers from multiple academic libraries, followed by deduplication. The deduplicated records are then passed to the second phase to evaluate the relevance of these candidate studies [5]. This is typically done in a double-blinded mode, meaning two reviewers screen all records to minimize human errors and bias. In case of conflicts, a third reviewer's opinion is used to resolve it [6]. Initially, researchers evaluate the relevance of each paper based on its title and abstract, comparing it to the already defined eligibility criteria. Records that meet all inclusion criteria and do not violate any of the exclusion criteria in this initial screening stage are subsequently subject to FT screening based on the same eligibility criteria. After the second screening stage, the appropriate data gets extracted from the remaining papers and included in a descriptive analysis and a flow diagram to ensure transparency and reproducibility [5].

The mean duration of an SR from the PROSPERO registry [7] is approximately 67 weeks [8], with TiAb and FT screening being the most time-critical phases. [9] analyzed 319 SR requests from the SR request data from Weill Cornell Medicine's service. Out of the 319 SR requests, 30% were abandoned during TiAb and 24% during FT screening, underscoring the criticality of these two screening stages.

Due to the remarkable performance improvements of LLMs across various downstream tasks over the last few years, several studies have experimented with automating screening in SRs using such models. By introducing Instruction Structure Optimized (ISO) prompting and their ISO-ScreenPrompt [10], researchers achieved results over 90% in terms of accuracy, sensitivity, and specificity on the training and validation datasets for FT screening. [11] demonstrated how Retrieval Augmented Generation (RAG) with GPT-4 [12] can effectively be used for FT screening. In their setting, the FT of each paper served as the document set from which the LLM retrieved information. The evaluation of their approach on one completed SR resulted in a specificity of 99.6%.

Other studies with either insufficient [13] or only in some cases on par with human [14] results underline the difficulty of automating FT screening with LLMs.

Given the limited amount of related work in FT screening, this study further focused on related work covering TiAb screening approaches. Studies were considered relevant if they were extensible and achieved a high sensitivity. Several studies on LLM-based TiAb screening have been found [15]–[17], but rarely any reach a sensitivity of greater than 99%. To be used in real-world practice as a replacement for human screeners, any automation approach must meet this sensitivity level, as required by Cochrane [18]. The 5-tier approach [3] is one study that introduced a scalable prompting strategy and reached the Cochrane sensitivity requirement. This was achieved by classifying papers into five categories, ranging from 1 (highly relevant) to 5 (not relevant). Papers that the LLM assigns to category 5 are excluded automatically, while those in categories 1 to 4 remain subject to human screening. This approach, which excludes only studies where the LLM is highly confident of ineligibility, maximizes sensitivity. However, the effectiveness of the 5-tier approach on open-source LLMs and its application to FT screening have not yet been investigated.

III. Methodology

Compared to the 5-Tier-Prompting case study [3], LLaMA 3.1 8B has served as LLM for evaluation instead of GPT-4. Figure 1 depicts the required steps to conduct experiments on TiAb and FT screening, which are subsequently described in detail. The code used to conduct the experiments can be found in the supplementary material provided through Zenodo [19].

As the SYNERGY dataset [20] was used for evaluation, the first stage of the pipeline included FT retrieval via the BioC API for PMC Open Access [21] followed by parsing the XML response. Due to restricted access to retrieve the FT of various papers, only the 5 largest SRs after retrieval were considered for evaluation as they most closely mirrored real-world conditions. The selected SRs had a substantial number of studies to screen with a relatively small proportion of studies ultimately included. In this paper, the term 'dataset' was used for each of the 5 selected SRs. Table I gives a brief overview of each dataset regarding the topic of the SR, number of total records, and number of records with decision 'include' as ground truth after retrieval.

TABLE I. DATASETS AFTER FT RETRIEVAL

Dataset	Topic(s)	Records	Included
Bos_2018	Medicine	1163	5
Brouwer_2019	Psychology, Medicine	6482	11
Leenaars_2020	Medicine	791	75
van_Dis_2020	Psychology, Medicine	1753	15
Walker_2018	Biology, Medicine	3234	88

In prompt construction, three different prompts have been used. The 5-Tier-Prompting approach [3] for TiAb screening served as the baseline, while two FT screening approaches have been introduced:

- **FT1**: The 5-Tier prompt has been adjusted accordingly by exchanging the terms 'title and abstract' to 'fulltext' (see Table II) and the FT has been passed instead of TiAb.
- **FT2**: In addition to the changes in FT1, the eligibility criteria have been positioned before and after the FT. This approach was inspired by the study of [10] to address the 'lost-in-the-middle' phenomenon [22]–[24].

Table III summarizes the prompt structure used in the three approaches. For the baseline (TiAb), the same approach as in [3] was used. The structure for FT1 was similar to that in [3], where the only adjustments were the new system prompt and

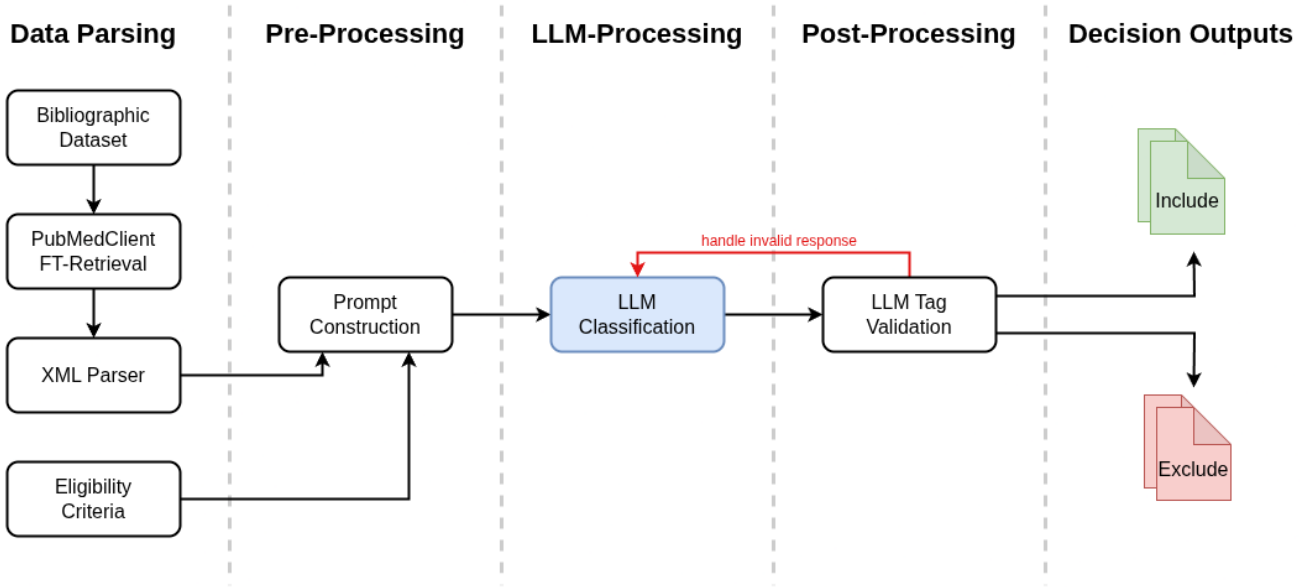


Figure 1. Conceptual Architecture.

TABLE II. SYSTEM PROMPT FOR FT SCREENING

Structure	Prompt
Role Play Instruction	You are a researcher rigorously screening full-texts of scientific papers for inclusion or exclusion in a review paper. Based on the provided inclusion and exclusion criteria listed below, you are asked to assign the paper to one of the following groups:
5-Tier-Group Definition	"1 Highly Relevant": Based on the given fulltext, the paper meets all inclusion criteria and no exclusion criteria. Therefore, the paper will be included. "2 Probably Relevant": The information provided in the full-text indicates that the paper is likely relevant. "3 Undecidable": The given full-text does not contain enough information to evaluate whether the inclusion and exclusion criteria are met. "4 Probably Irrelevant": Based on the given full-text, it is likely that at least one inclusion criterion is not met or that at least one of the exclusion criteria is met. "5 Not Relevant": Based on the full-text, it is clear that the paper does not meet the criteria. Therefore, the paper will be excluded.
Response Instruction	Based on the probability of a paper meeting all inclusion criteria and no exclusion criteria, assign it to one of the five categories. Only type the number of the group as "1", "2", "3", "4" or "5" in your response. Do not type anything else.

TABLE III. PROMPT STRUCTURE FOR SCREENING

TiAb	FT1	FT2
System Prompt [3]	System Prompt Table II	System Prompt Table II
Title and Abstract	Full-Text	Eligibility Criteria
Eligibility Critera	Eligibility Criteria	Full-Text
		Eligibility Criteria

the provision of FT instead of TiAb. FT2 is an extension of FT1, which aimed to address the 'lost-in-the-middle' phenomenon.

The constructed prompt was then passed to a locally hosted LLaMA 3.1 8B model for screening. By including a validation function to check whether the response of the LLM was only a number between 1 and 5 (as requested in the prompt), invalid responses were eliminated. If, after five retries, the response for a paper was still invalid, the paper got manually assigned '1' as tag value. By assigning '1', the paper was subject to human screening in the setting, as the LLM was not able to provide a valid screening decision.

This modular architecture allows the approach to be adapted to alternative FT retrieval mechanisms and other LLMs. Additionally, the prompt construction is independent of the underlying eligibility criteria and candidate studies, enabling applicability to large-scale systematic reviews without restriction to specific topics or domains.

Similar to [3], evaluation is based on two metrics: sensitivity and workload reduction. Sensitivity as defined in (1) is the fraction of ground-truth includes classified by the LLM as include. It measures the risk of missing relevant studies. Workload reduction (2) is based on the assumption that screening automation is integrated into the systematic review workflow as a filtration step, whereby records classified as exclude are removed prior to the human screening phase. Consequently, it represents the proportion of papers excluded by the model.

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive + False Negative}} \quad (1)$$

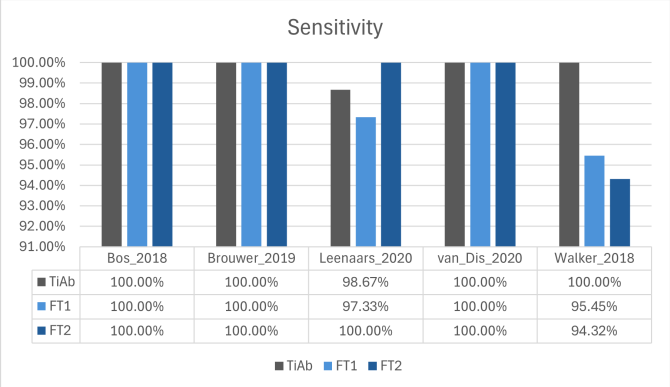
$$\text{Workload reduction} = \frac{\text{True Negative + False Negative}}{\text{n}} \quad (2)$$

where n represents the total number of papers

IV. RESULTS

As in [3], alpha = 4 was used to transform the LLM response into a binary classification for further analysis, meaning that all

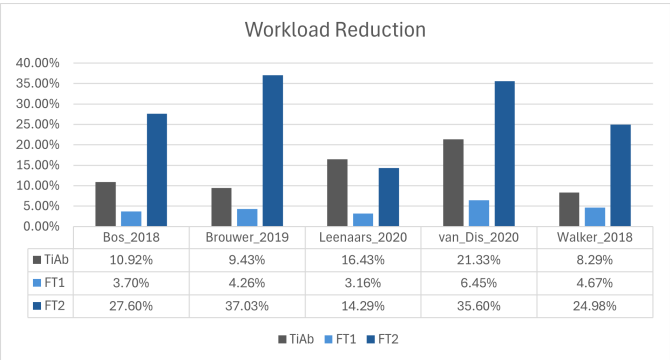
records with a tag value smaller or equal to 4 got the decision 'include' and entries with tag value 5 got the decision 'exclude'. Figure 2 shows the sensitivity per dataset for each prompting approach. FT1 was on par with TiAb on three datasets, while FT2 outperformed TiAb in Leenaars_2020, where TiAb had 98.67% and FT 100% sensitivity. Overall, in only 4 cases, a sensitivity of less than 100% was achieved. For Walkers_2018 both FT approaches and Leenaars_2020 TiAb and FT1, the sensitivity was less than 100%. The lowest sensitivity score had FT2 for Walkers_2018 with only 94.32%.



	Bos_2018	Brouwer_2019	Leenaars_2020	van_Dis_2020	Walker_2018
TiAb	100.00%	100.00%	98.67%	100.00%	100.00%
FT1	100.00%	100.00%	97.33%	100.00%	95.45%
FT2	100.00%	100.00%	100.00%	100.00%	94.32%

Figure 2. Sensitivity per Dataset for each Prompting Approach.

In Figure 3, the workload reduction per dataset is visualized. FT2 outperformed TiAb in 4 out of 5 datasets, whereas FT1 showed the weakest workload reduction of all three approaches in every dataset. In the most extensive dataset, Brouwer_2019 with 6482 records, FT2 outperformed TiAb in terms of workload reduction, achieving 37.03% compared to 9.43%, whereas TiAb outperformed FT2 on the smallest dataset, Leenaars_2020 with 791 records, with 16.43% compared to 14.29%.



	Bos_2018	Brouwer_2019	Leenaars_2020	van_Dis_2020	Walker_2018
TiAb	10.92%	9.43%	16.43%	21.33%	8.29%
FT1	3.70%	4.26%	3.16%	6.45%	4.67%
FT2	27.60%	37.03%	14.29%	35.60%	24.98%

Figure 3. Workload Reduction per Dataset for each Prompting Approach.

Figure 4 and Figure 5 give valuable information when comparing the three different approaches by visualizing the weighted averages of sensitivity and workload reduction for TiAb, FT1, and FT2. Given the varying sizes of the datasets used in the evaluation, the results have been weighted by the size of each dataset to obtain weighted averages for sensitivity and workload reduction. In this way, the weighting prevents the largest dataset from disproportionately influencing the average performance of each screening method.

On average, TiAb screening achieved a sensitivity of 99.73%, while FT1 and FT2 had 98.56% and 98.86% respectively.

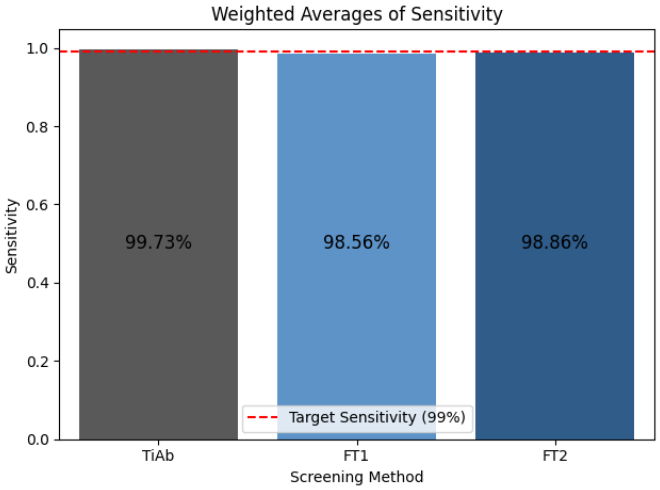


Figure 4. Weighted Average of Sensitivity for each Prompting Approach.

However, when comparing the workload reduction of each approach, FT2 achieved the highest result with 27.9%, almost 15% higher than TiAb, which had a 13.28% reduction. FT1 turned out to perform weakest in terms of workload reduction with only 4.45%.

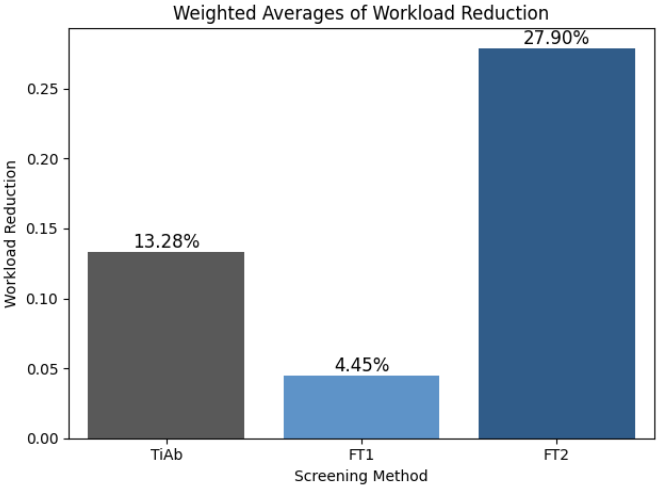


Figure 5. Weighted Average of Workload Reduction for each Prompting Approach.

V. Discussion

The results showed that when using LLaMA 3.1 8B, only LLM-powered TiAb screening meets Cochrane's sensitivity requirement (sensitivity $\geq 99\%$).

The experiments using FT were based on the assumption that additional context might increase the workload reduction by reducing the number of irrelevant papers included by the LLM, while maintaining a high sensitivity.

FT1 turned out to be the least favorable setting, achieving the lowest sensitivity and workload reduction among all three approaches. The results were counterintuitive to the assumption that additional context might help the LLM to make better screening decisions. Hence, FT2 was introduced to check whether the task was too complex for the model or, due to the long context of the instruction and paper, the LLM got lost in the middle of processing and 'forgot' about eligibility criteria. The only difference between FT1 and FT2 was the positioning of the eligibility criteria. By setting the eligibility criteria before and after the FT in the prompt, sensitivity increased slightly, while workload reduction improved significantly. [10] also reported performance increases when positioning criteria before and after FT to address the 'lost-in-the-middle' phenomenon, confirming the plausibility of the FT2 approach.

The lower sensitivities of the FT approaches compared to TiAb mean that the LLM wrongly classified certain papers as 'exclude' while the ground truth was 'include'. When comparing papers against the eligibility criteria, humans naturally focus more on methodological chapters and results. When providing the FT of a paper to an LLM, chapters such as the related work and future work might be misleading and could influence the screening decision.

VI. Conclusion and Future Work

The conducted experiments gave valuable insights into LLM-powered TiAb and FT screening. Additional context, provided by passing the FT instead of only the TiAb, showed no further improvements in terms of sensitivity. Other approaches need to be considered to verify whether additional context might help to constantly reach over 99% of sensitivity. However, the increase in workload reduction in the FT2 setting indicates that additional context provided enhances the decision making of LLMs during screening.

A more extensive evaluation dataset, with SRs from different topics, could confirm the robustness of the 5-Tier-Prompting approach for FT screening. Given the rapid evolution of LLMs, an evaluation with newer models could provide further insights. As this study focused on the use of an open-weight model, a comparison of results between LLaMA 3.1 8B and newer LLaMA models could give further insights into whether the currently false negative classifications are occurring due to the high complexity of the task. Lastly, not all parts of a paper are likely to be relevant during the screening process. New experiments, where LLMs are enhanced to focus more on relevant chapters by either changing the current prompt or introducing pre-processing of FTs, might further improve performance.

In summary, this study confirmed the significant potential of the 5-Tier-Prompting approach. Although the extension of the approach by considering the FTs requires further evaluation, the results with LLaMA 3.1 8B are promising and potentially open up even better results with newer open-weight models. Nonetheless the small decrease in terms of sensitivity when using FT needs to be further investigated.

Acknowledgements

The joint CERN and WHO ARIA [25] project is funding the PhD project of Elias Sandner and the research project and stay at CERN of Marko Zeba.

References

[1] B. Nussbaumer-Streit *et al.*, "Resource use during systematic review production varies widely: A scoping review", *Journal of clinical epidemiology*, vol. 139, pp. 287–296, 2021.

[2] K. G. Shojania *et al.*, "How quickly do systematic reviews go out of date? a survival analysis", *Annals of internal medicine*, vol. 147, no. 4, pp. 224–233, 2007.

[3] E. Sandner *et al.*, "Screening Automation for Systematic Reviews: A 5-Tier Prompting Approach Meeting Cochrane's Sensitivity Requirement", in *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, Dubai, United Arab Emirates: IEEE, Nov. 2024, pp. 150–159, ISBN: 979-8-3503-5479-9. DOI: 10.1109/FLLM63129.2024.10852425.

[4] A. Pollock and E. Berge, "How to do a systematic review", *International Journal of Stroke*, vol. 13, no. 2, pp. 138–156, 2018, PMID: 29148960. DOI: 10.1177/1747493017743796. eprint: https://doi.org/10.1177/1747493017743796.

[5] E. Calderon Martinez *et al.*, "Ten Steps to Conduct a Systematic Review", *Cureus*, vol. 15, no. 12, e51422, 2023, ISSN: 2168-8184. DOI: 10.7759/cureus.51422.

[6] J. P. Higgins *et al.*, *Cochrane Handbook for Systematic Reviews of Interventions Version 6.5 (updated August 2024)*. Cochrane, 2024.

[7] Centre for Reviews and Dissemination, University of York, "Prospero: International prospective register of systematic reviews", 2025, [Online]. Available: https://www.crd.york.ac.uk/prospero/ (visited on 09/05/2025).

[8] R. Borah, A. W. Brown, P. L. Capers, and K. A. Kaiser, "Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry", eng, *BMJ open*, vol. 7, no. 2, e012545, Feb. 2017, ISSN: 2044-6055. DOI: 10.1136/bmjopen-2016-012545.

[9] M. R. Demetres, D. N. Wright, A. Hickner, C. Jedlicka, and D. Delgado, "A decade of systematic reviews: An assessment of Weill Cornell Medicine's systematic review service", *Journal of the Medical Library Association : JMLA*, vol. 111, no. 3, pp. 728–732, Jul. 2023, ISSN: 1536-5050. DOI: 10.5195/jmla.2023.1628.

[10] C. Cao *et al.*, *Prompting is all you need: LLMs for systematic review screening*, Jun. 2024. DOI: 10.1101/2024.06.01.24308323.

[11] F. Trad *et al.*, *Streamlining Systematic Reviews: A Novel Application of Large Language Models*, Dec. 2024. DOI: 10.48550/arXiv.2412.15247. arXiv: 2412.15247 [cs].

[12] OpenAI, "Gpt-4 system card", 2023, [Online]. Available: https://cdn.openai.com/papers/gpt-4-system-card.pdf (visited on 09/05/2025).

[13] X. Chen and X. Zhang, *Large language models streamline automated systematic review: A preliminary study*, 2025. arXiv: 2502.15702 [cs.IR].

[14] Q. Khraisha, S. Put, J. Kappenberg, A. Warraitch, and K. Hadfield, "Can large language models replace humans in systematic reviews? Evaluating GPT -4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages", *Research Synthesis Methods*, vol. 15, no. 4, pp. 616–626, Jul. 2024, ISSN: 1759-2879, 1759-2887. DOI: 10.1002/jrsm.1715.

[15] A. Huotala, M. Kuutila, P. Ralph, and M. Mäntylä, *The Promise and Challenges of Using LLMs to Accelerate the Screening Process of Systematic Reviews*, May 2024. DOI: 10.48550/arXiv.2404.15667. arXiv: 2404.15667 [cs].

[16] F. Dennstädt, J. Zink, P. M. Putora, J. Hastings, and N. Cihoric, "Title and abstract screening for literature reviews using large language models: An exploratory study in the biomedical domain", *Systematic Reviews*, vol. 13, no. 1, p. 158, Jun. 2024, ISSN: 2046-4053. DOI: 10.1186/s13643-024-02575-4.

[17] L. Affengruber *et al.*, "An exploration of available methods and tools to improve the efficiency of systematic review production: a scoping review", *BMC Medical Research Methodology*, vol. 24, no. 1, p. 210, Sep. 2024, ISSN: 1471-2288. DOI: 10.1186/s12874-024-02320-4.

[18] J. Thomas *et al.*, "Machine learning reduced workload with minimal risk of missing studies: Development and evaluation of a randomized controlled trial classifier for cochrane reviews", *Journal of Clinical Epidemiology*, vol. 133, pp. 140–151, 2021.

[19] E. Sandner *et al.*, *From abstracts to full texts: The impact of context positioning in llm-based screening automation*, Accessed: 2025-10-21, 2025. DOI: 10.5281/zenodo.16419874.

[20] J. De Bruin, Y. Ma, G. Ferdinands, J. Teijema, and R. Van de Schoot, *SYNERGY - Open machine learning dataset on study selection in systematic reviews*, version V1, 2023. DOI: 10.34894/HE6NAQ.

[21] National Center for Biotechnology Information, "Bioc-pmc: Biomedical natural language processing apis", 2025, [Online]. Available: https://www.ncbi.nlm.nih.gov/research/bionlp/APIs/BioC-PMC/ (visited on 09/05/2025).

[22] S. An *et al.*, "Make your llm fully utilize the context", *Advances in Neural Information Processing Systems*, vol. 37, pp. 62 160–62 188, 2024.

[23] N. F. Liu *et al.*, "Lost in the middle: How language models use long contexts", *arXiv preprint arXiv:2307.03172*, 2023.

[24] G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis, "Efficient streaming language models with attention sinks", *arXiv preprint arXiv:2309.17453*, 2023.

[25] World Health Organization, "Aria tool on who partners platform", 2025, [Online]. Available: https://partnersplatform.who.int/tools/aria (visited on 09/05/2025).

Electronic Health Records and the Archival Question: Shared Responsibility as a Panacea

Mehluli Masuku
History Department. Sorbonne University Abu Dhabi
SAFIR, Centre for Humanities, Languages and Education, United Arab Emirates
e-mail: mehluli.masuku@sorbonne.ae

Abstract— **Literature on the benefits of Electronic Health Records (EHR) across the health sector abounds. However, as all the hype about EHRs revolutionizing healthcare, there is not a commensurate discourse on the archival question in the era of EHRs. As EHRs are being hailed and hyped, an archival dilemma of such systems is looming. The archival question is about the long-term availability of EHRs as archives versus the traditional custodial roles of archival institutions. Through an argumentative approach, the paper argues that due to their complexities, EHRs need not be transferred to archival institutions but rather require a distributed custody strategy where archival institutions allow healthcare facilities and vendors to retain and preserve EHRs to ensure that their EHR systems guarantee long term preservation of records generated by such systems (as archivally deemed fit), and that healthcare facilities implement and maintain archivally compliant EHRs archives. In the process, the distributed custody will ensure that archival institutions promote the availability of the record for now and the future by collaborating with healthcare facilities as an archival protector whilst the custody of the EHR remains with the healthcare provider. For this approach to work, archival laws need to be reviewed in order to reflect and cope with the new realities and address current challenges presented by EHRs. It is suggested that archivists, health IT experts and vendors start viewing EHRs not merely as systems from which archival health records will be extracted at a later stage, but as health records in their own right, that need to be treated as archives, as they are resident in the systems.**

Keywords-archiving; electronic health records; long term preservation; shared responsibility.

I. INTRODUCTION

EHRs are fast becoming ubiquitous across the health sector worldwide. Ironically, there seems to be a dearth of commensurate discourse on the archival question around EHRs. The situation of limited attention being given to the archiving of clinical data persists to-date in the EHRs [1]. An EHR is "medical information compiled in a data-gathering format for retention and transferal of protected information via a secured, encrypted communication line. The information can be readily stored onto an acceptable storage medium, such as a compact disk" [2]. They have been hailed for revolutionizing the health sector by overcoming the shortcomings of their predecessors—paper records, including illegibility of handwriting, limited shareability, as well as space challenges. With the advent of Artificial Intelligence (AI), Machine Learning (ML), big data analytics and many other forms of computing, the value of EHRs has only increased. The advent of Electronic Personal Health Records (EPHR) where patients have control over their own health records has accelerated the interest of researchers in EHRs. For example, some scholars are beginning to advocate for EHRs preservation models that will see the control of EHRs being moved away from healthcare facilities to their owners [3]. The question of EHRs ownership is another one that is a highly contested issue, but not a topic for discussion in this paper. However, some scholars are already acknowledging that EHRs are currently residing with healthcare providers, not with archival institutions, further demonstrating the need for a genuine debate about EHRs preservation and the custody question [3]. This indirect acknowledgement by certain scholars shows that healthcare facilities with EHRs vendors, have become the default archivists of EHRs.

The phenomenon of EHRs has been, and continues to be, a subject of research in health informatics for several years now [4]. However, as all this is happening, archivists are conspicuously absent from the stage regarding how EHRs should be archived for long term availability as archives. Such absence of archivists and archivally sound techniques and policies has not only created an archival chasm but has seen the question of long-term availability of EHRs—which is usually discussed with the ownership of such records, as a matter between healthcare facilities and EHRs vendors. What France et al. [5] identified as the challenge for the long-term preservation of EHRs still stands today and continues to make the dream of EHRs archiving elusive to archivists and all the other stakeholders. It can be observed that EHRs and most information systems for hospitals and all the other healthcare facilities, private and public, are built considering the "active" EHR [5]. This implies that the archival function for EHRs comes as an afterthought, if ever it arises, too late a stage for archival institutions to ingest the EHRs for posterity, owing to large volumes, multiple vendors, ever-changing technical standards etc., coupled with the freedom of healthcare facilities to select, implement and discontinue EHRs at their own discretion. This further results in a state where health records generated and held by EHRs are not subjected to archival processes, such as

appraisal and preservation, with the consequence of failure to identify and preserve archivally valuable health records. Owing to the dearth of clearly pronounced archival policy and a review of archival laws as they relate to the archival question for EHRs, archival institutions, especially national ones that are legally mandated to acquire and preserve such records, find themselves in a position where they are not able to execute their mandate, or sub consciously abrogating their archival role and duty to healthcare providers and EHRs vendors. Writing in the context of the USA when EHRs were still a relatively new and buzzy phenomenon, Corn warned that "several ongoing health care and information storage developments suggest that a fresh look should be taken at the policies that govern the preservation of medical records…" [1]. One can say that the archival question in terms of the legal mandate or responsibility versus practical considerations for EHRs preservation is part of Corn's statement.

The current state of literature shows preoccupation with privacy and security of health data in EHRs, but little has been said about the real archival question, that is, are archival institutions abrogating their legal responsibilities of acquiring and preserving health records to healthcare facilities and vendors without explicitly mentioning this? Corn, referring to the question about the preservation of clinical data as a rarely asked question, further stated *"how to archive clinical data so as to preserve most efficiently the information and access to the information will require collegial resolution of complex technical and social problems, analogous to the efforts of many major libraries to develop archiving policies. Reevaluation of U.S. medical record retention policies would require consideration of a number of issues that are not often discussed" [1].*

This paper is structured as follows. In Section I, an introduction to the study is given. In Section II, a case for the preservation of EHRs is presented. In Section III, the archival challenges of preserving EHRs are given. In Section IV, the shared responsibility approach to EHRs is proposed. Finally, the study is concluded in Section V.

II. WHY SHOULD HEALTH RECORDS BE ARCHIVED?

The importance of medical archives cannot be overemphasized. According to the Mansoura University Specialized Medical Hospital, a medical archive is a "fundamental pillar in any health institution, as it is the memory that preserves the history of patients and their health conditions, as it is responsible for preserving the services provided to patients, whether in paper or electronic form, for reference when needed" [6]. Medical archives support quality healthcare and evidence-based decision making, advance medical research and protect healthcare facilities from legal liabilities [6]. From a global perspective, "national laws state that medical record should be preserved for patient's health care and for legal purposes, for a certain number of years, after his departure or after the patient's death. This duration varies by country" [5]. Although medical records are primarily created for the care of patients, medical archives are used for a number of reasons, including medical practitioners understanding previous case history in the event of the case resurfacing, historians studying the history of medicine, fiction authors wishing to understand a contemporary setting or character for their stories as well as genealogists researching their relatives [and other persons of their interest] [7]. Despite the acknowledgement of the secondary use of health records, progress towards their preservation for such meaningful use remains a serious hurdle. In the following words, Klementi and colleagues, lament: "although the need for health data reuse is widely recognised, actual progress in that area has been moderate. The reasons for this are the vendor-specific proprietary database schemes used by EHRs systems, semantic heterogeneity and the sensitive nature of clinical data that sets legal and ethical restrictions on sharing" [3].

The study was guided by the following objectives.

1. To unpack the archival challenges that are presented by EHRs.
2. To propose the distributed custody approach for the preservation of EHRs.

III. ARCHIVAL CHALLENGES PRESENTED BY EHRS

The complexity of EHRs architecture is the culprit, presenting a domino effect on the other aspects, such as the ability of archival institutions to cope with technical and custodial challenges of such systems. EHRs are not just a replacement of paper records. The complexity of EHRs "resides in a multitude of interdependent elements which must be organized" [4]. Franca, Lima and Soares, summarising the complexities of EHRs, wrote: "*One health application that is considered very complex not only to develop but also to operate and maintain is the Electronic Health Record (EHR), which refers to software systems that store health information about patients in a digital format. EHRs are used by different health professional teams, including physicians, nurses, radiologists, pharmacists, laboratory technicians and radiographers. Even patients can add information into the EHR, provided this is validated by physicians" [8].*

Thus, EHRs are complex integrated systems from various operational units and functions of health. This makes it difficult for national archival institutions to play their traditional roles of acquiring, managing and archiving EHRs. This complexity further implies that EHRs are fluid and ever active, which further complicates the archival work of identifying what constitutes archivally worth records in such systems. A hospital archive consists of both active health and passive health records [9]. Active health records, on one hand, include those whose data are being updated on a continuous basis of a patient receiving care at a healthcare facility, whilst passive health records, on the other hand, are health records that have not been updated over a reasonably long period of time [9].

In today's EHRs environment, determining passive health records is increasingly becoming difficult due to the treatment of EHRs — they are ever active data used in big data analytics, training AI algorithms and in ML. Further compounding the situation is how to archive EHRs themselves, that is, should the entire EHR system be archived, or only selected components of the EHR should be archived? Some of the early scholars, such as France and colleagues suggested that certain records be extracted from EHRs systems and earmarked for archiving [5]. They wrote "data will have to be extracted from the administrative database (admission, dates of admission and discharge, clinical services, notes of the physician, nurses, drug orders, laboratory computers, X rays, surgery and anaesthesia protocols, final report, medical record summary...)" [5].

However, they immediately lamented: "Will the record be still accessible? Will it be complete ?" [5]. For example, in their study in the long term preservation abilities of EHRs among healthcare facilities in Barcelona, Bote, Termens, and Gelabert discovered that while most analogue health records easily became passive within a period of 3 – 5 years, the case was likey to be different from EHRs systems as information workflows and treatment in EHRs are completely different from an analogue environment and both active and passive EHRs were lying in the same electronic systems that however, were not running on the same software [9]. So, how do archival institutions exercise their archival role in such complex environments? This is a critical question for archivists as their response to it will determine how to perform the appraisal function for EHRs.

The question is, can archival institutions cope with these EHRs complexities from the perspective of physical custodianship and guaranteeing long term availability and accessibility of the EHRs? For example, if the entire EHR is to be archived, the question is who should do this and how? Should healthcare facilities migrate their EHRs system once they become due for archiving? Should the archival institution ingest the EHR system or migrate it to a new system for long term preservation? Bote, Termens, and Gelabert advise that an archival shift is necessary in such an environment [9]. Therefore "an EHR is a complex unit of information that requires different treatment in the long term, not as a single unit of information." [9].

When the headache of long-term preservation of EHRs storage started, some attempts were made, for example, The Swedish Institute for Health Services Development, working with Sahlgrenska University Hospital created the project called DEJAVU, whose aim was to come up with a general format for the long-term preservation of electronic patient records [10]. At that time, the Swedish National Archives had approved paper and microfilm as the most stable and permanent storage mediums for archival preservation [10]. The project proposed the development of a standardised electronic patient record based on the Standard Generalized Markup Language, ISO 8879:1986/A1 :1988, allowing records to be extracted from their original EHRs systems and converted into a Standard Generalised Markup Language (SGML) format for long-term accessibility [10]. However, the susstainability of such a pactice, as well as the ability of the new and stable SGML EHR to remain interactive and comprehensive over time under the care of an archival institution remains a concern, given that records would have to be extracted from their original systems into the SGML. To date, a plethora of EHRs standards have been developed, including standards from Health Level Seven (HL7), International Organisation for Standardisation (ISO)/ Technical Committee (TC) 215 Secretariat on Health informatics. Some of the relevant standards by ISO include 1SO 14641:2028 Electronic Document management- Design and operation of an information system for the preservation of electronic documents- Specifications, the European Committee for Standardisation (CEN) 254 and many others. However, the afore-mentioned standards are best applied to the EHRs systems that create, store and preserve such records, rather than applying them to third parties, such as archival institutions that will have to extract such records from their original EHRs, covert them to new formats for the sake of long term preservation. Questions of records completeness in the midst of a multiplicity of EHRs components and EHRs vendors arise.

Limited collaboration between health records and IT personnel at healthcare facilities and the national archival institutions also have a domino effect on the ability of archival institutions to preserve EHRs. As Corn put it "another important complicating issue for the archiving of electronic information is the fragmentation of responsibility and diversity of technical systems for varying data types within a single enterprise: laboratory, pathology, imaging, sensor information, and other forms of data are often maintained by the unit performing the service, while the EHR proper is the responsibility of the organization's IT department." [1]. As the above mentioned situation continues to obtain across the EHR environment, it is obvious that the bulk, if not all the health records—active or inactive (archival), generated by EHRs is under the control of IT staff at healthcare facilities and vendors, thereby making them "default archivists", yet there is limited dialogue between archivists and such technical personnel who are in real control of health records. Even in cases where there is a medical record librarian or archivist at a healthcare facility, they tend to bear influence on traditional paper records only.

Thus, the fragmentation is not only witnessed in the components of the EHR system, but between and among the personnel and institutions handling EHRs across the spectrum, including archivists and Archives, IT personnel in healthcare facilities and vendors. This has resulted in disjointed and limited efforts towards addressing the archival question as far as EHRs are concerned. This has manifested itself in the form of policies that are focused on the regulation of operational issues of EHRs, with a loud

silence about the archival question, especially in terms of whose responsibility it is to perserve the archival health record beyond its operational value for immediate care of a patient. Fig 1 summarises the challenges associated with the preservation of EHRs from an archival perspective.

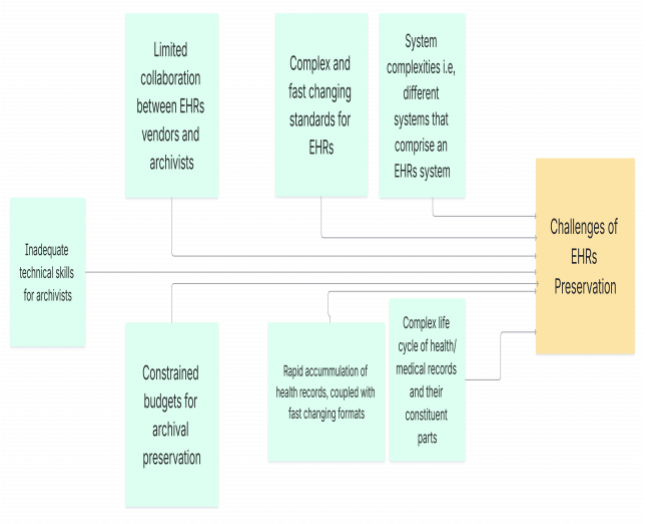


Fig. 1 Challenges of preserving EHRs.

IV. SHARED RESPONSIBILITY AS A MEANS OF ADDRESSING THE ARCHIVAL QUESTION FOR EHRS

As noted by Bastian [11], the idea of shared responsibility in archival science, also known as post-custodialism, was first coined by Ham, who encouraged archivists to rethink their physical custodial role over archives and allow a situation where archivists and archives let of the physical custody of records whilst retaining their legal control. This implies that the creating agency is left to maintain physical custody of the records, including archival ones, with archivists coming in to advise on best archival practices. To Ham, cited in Bastian [11], the insistence of archivists on the physical control of records, especially in the era of electronic records, is seen as a deterrent to the effective management of records.

Despite the shared responsibility having a relatively long history in archival science, it is yet to gain scholarly and practical traction in archival science as a handful of scholars have researched it. In light of EHRs, Ham's post-custodial approach is being proposed as the most plausible one to the preservation of EHRs. Whilst early scholars recommended a transfer of EHRs to standard archival storage medium, such as microfilm, DVD and CD-ROMS, clearly this system is no longer viable today. Despite such medium having become an obsolete technology, the very approach of extracting records from EHRs systems into a different platform implies migration from one platform or medium to the other, which in the case of EHRs, I argue, is not viable, due to the complications of the EHRs today and in the near and long future. The multiplicity of different systems for different health operations during the lifetime care for a patient, supported by different vendors, operating on different platforms, technical incapacities of many archival institutions as well as incommensurate archival strategies for EHRs all work against the traditional custodial approach to the long-term preservation of EHRs. As a result, the post-custodial/shared responsibility model, which archivists have long touted but not yet supported by clear archival policies and laws due to the fear of entrusting the archival function to healthcare facilities and vendors, is proposed.

In this model, archival institutions and archivists stick to their legal supervisory mandate and loosen up on the custodial component. This means that archival institutions focus on the development of policies and advocate for the updating of archival laws that promote the shared responsibility model. Specifically, archivists may stop insisting on taking custody of EHRs as this demands them to have as many software applications as the number of EHRs variations that exist on the market, which is neither practical nor economic for poorly funded archival institutions. This will see archivists being responsible for setting policies and standards that EHRs vendors and healthcare facilities should ensure as they share the custody of EHRs, including retention and disposal schedules as well as access policies for EHRs.

Likewise, healthcare facilities and vendors where necessary, will be expected to take full or shared custody of EHRs systems, embedding in them, the archival policies and standards for the records to meet the archival needs of the health record, including research uses. This means that EHRs will no longer follow the traditional custodial cycle of records where in-active records are handed over to the archives, but remain in the system of the healthcare facility, in conjunction with vendors where necessary, to preserve them as archives. This means allowing the healthcare facilities and vendors to protect and make available, legally acceptable EHRs for as long as required by the law under the custody of their creators, duties that traditionally lie squarely with archivists and archival institutions. In this model, archivists, IT specialists and EHRs vendors work together and come up with the best framework that factors in the technical, operational and legal realities of all the stakeholders, at the same time protecting and propagating EHRs that meet legal and archival requirements for health records. Through this model, the policy, technical and communication fragmentation that currently characterizes the EHRs can be reduced, and a more coordinated long-term EHRs can be imagined.

The proposed distributed custody model, however, will not come without challenges. Whilst from a custodial perspective, this approach seems to be a panacea, it creates a new demand from archivists and archival institutions—that of educating and ensuring compliance with archival policies and standards for EHRs vendors and healthcare facilities. The success of the model depends on the cooperation of the vendors and healthcare facilities. The archival endeavor has

been the prerogative of archivists, with creators and vendors only interested in the technology that supports the daily operations of healthcare facilities, with little to no regard for the records when they have fulfilled their primary administrative needs. Thus, giving them the archival role of preserving their in-active EHRs may not enjoy positive reception. On the part of the healthcare facilities and vendors, the distributed custody model may come with additional expenses, for example, maintaining archival EHRs accessible as archivists would diligently do. Archival preservation is an expensive and cumbersome endeavor that vendors and healthcare facilities may not keep pace with. Thus, the model presents new responsibilities to archivists, healthcare facilities and the vendors, requiring a high degree of commitment and consensus from the three stakeholders.

V. CONCLUSION AND FUTURE WORK

Although archivists have been reluctant to adopt and promote the shared responsibility approach to address the archival question in the case of electronic records, owing to them being "collecting professionals" with the keeping of records in the archives as a way of vouching for their authenticity, EHRs coupled with their complexities, undoubtedly present themselves as a good candidate for this approach. The proposed shared responsibility approach that this paper advocates for is going to be challenging as it needs very strong coordination efforts among archivists, healthcare IT experts and vendors, as well as a great deal of a mindset shift from the afore-mentioned stakeholders, allowing each of them to adapt to their new professional roles in this model, including incurring expenses.

It is recommended that archivists, EHRs vendors and healthcare organisations engage in feasibility discussions on the distributed custody approach for EHRs. Archivists should develop policy and compliance mechanisms to ensure the success of the model, whilst vendors and healthcare providers should be willing to bear the archival responsibility of ensuring the long-term availability of EHRs in their custody, at the same time fulfilling the information needs of society and researchers. Empirical studies are required to understand the feasibility of the distributed custody approach from policy, legal, technical and financial considerations for the model to work.

REFERENCES

[1] M. Corn, "Archiving the phenome: clinical records deserve long-term Ppeservation," *Journal of the American Medical Informatics Association*, vol. 6, pp. 1– 6, 2009. DOI: doi: 10.1197/jamia.M2925.

[2] W. Bartschat et al., "The legal process and electronic health records," *Journal of AHIMA*, vol. 76, pp. 96A-D, 2005.

[3] T. Klementi, K. J. I. Kankainen, G. Piho, and P. Ross, "Prospective research topics towards preserving electronic health records in decentralised content-addressable storage networks," The International Health Data Workshop (HEDA-2022), Jun. 2022, CEUR Workshop Proceedings. [Online]. Available from: https://ceur-ws.org/Vol-3264/HEDA22_paper_7.pdf

[4] L. Dobrica, C. Alexandra, and C. T. Ionescu, "Towards a better understanding of EHR systems using architectural views," Proc. International Conference on Health Informatics (HEALTHINF-2013), pp. 362-365, ISBN: 978-989-8565-37-2, DOI: 10.5220/0004246603620365.

[5] F. H. France, C. Beguin, R. van Breugel, and C. Piret, "Long term preservation of electronic health records: recommendations in a large teaching hospital in Belgium," Stud. Health Technol. Inform., pp. 632-6, 2000.

[6] Mansoura University Specialised Medical Hospital, "Medical archive.", 2025.

[7] Archives of Ontario, "Medical records at the Archives of Ontario: an overview.", 2012-25.

[8] J. M. S. Franc¸, J. de S. Lima, and M. S. Soares, "Development of an electronic health record application using a multiple view service oriented architecture," Proc. of the 19th International Conference on Enterprise Information Systems (ICEIS 2017), vol. 2, pp. 308-315, 2017. DOI: 10.5220/0006301203080315.

[9] J. Bote, M. Termens, and G. Gelabert, "Evaluation of healthcare institutions for long-term preservation of electronic health records", In. M.M. Cruz-Cunha et al. (Eds.): CENTERIS 2011, Part III, CCIS 221, pp. 136–145, 2011.

[10] T. Wigefeldt, S. Larnholt, and H. Peterson, "Development of a standardized format for archiving and exchange of electronic patient records in Sweden," Stud. Health Technol. Inform., vol. 43, pp. 52-56, 1997.

[11] J. A. Bastian, "Taking custody, giving access: a postcustodial role for a new century," Archivaria, vol. 53, pp. 76-93, 2004.

Using Data and Artificial Intelligence to Enable Successful Hospital at Home Programs

James P. Mcglothlin
RSM US LLP
Dallas, TX USA
jamie.mcglothlin@rsmus.com

***Abstract*—During the COVID-19 pandemic, health systems and payers had to take novel and extraordinary approaches to create hospital capacity and avoid hospital infections. Hospital at home programs have existed for years, but the pandemic environment led to additional interest, funding and reimbursement approvals for these programs. The hospital at home program is simply the monitoring and treatment of an acute inpatient patient at their residence. Research shows that outcomes and patient experience can be better with a hospital at home stay, while costs are much less. Nonetheless, these programs are relatively new and there are not well-researched standards for choosing patient cohorts or monitoring patient progress. Most hospitals either have a very restricted definition of eligible patients or leave the recommendation to the attending physician. While there are many case studies around the success of individual patients in hospital at home programs, there has been little research into choosing patient cohorts. In this study, we propose to use existing research around clinical trial enrollment and clinical data mining to better identify patient cohorts. We propose to use existing predictive analytics solutions to predict outcomes, adverse events and resource needs for individual patients. By combining all of these approaches, we can identify patients who are likely to succeed with hospital at home treatment, and we can monitor these patients and intervene to avoid risk of complications or adverse events.**

Keywords-hospital at home; care pathways; quality; artificial intelligence; predictive analytics; supervised learning; data mining; cardiology.

I. INTRODUCTION

The Hospital at Home (HaH) model is an innovative approach to healthcare delivery that provides acute, hospital-level care to patients in their homes. First conceptualized in the 1990s by researchers at Johns Hopkins University, HaH was developed to address rising healthcare costs, hospital overcrowding, and the desire to improve patient-centered care [1]. The model includes comprehensive medical services such as intravenous medications, oxygen therapy, diagnostic imaging, and frequent clinical monitoring, traditionally available only in inpatient settings. Evidence has demonstrated that HaH can be safe and effective for managing conditions, such as heart failure, Chronic Obstructive Pulmonary Disease (COPD), and community-acquired pneumonia, with outcomes that are equivalent or superior to traditional hospitalization in terms of mortality, readmission rates, and patient satisfaction [2] [3].

The COVID-19 pandemic significantly accelerated the adoption of HaH programs globally. In the United States, the Centers for Medicare & Medicaid Services (CMS) introduced the Acute Hospital Care at Home waiver in 2020, allowing eligible hospitals to receive full Medicare reimbursement for delivering inpatient-level services at home [4]. Since then, hundreds of health systems have implemented or expanded HaH initiatives, citing benefits such as reduced exposure to hospital-acquired infections, shorter lengths of stay, and improved patient experience [3]. Despite these advances, challenges persist, including regulatory uncertainty, reimbursement limitations in non-pandemic contexts, provider hesitancy, and the need for reliable home infrastructure and caregiver support [5]. As healthcare systems aim to modernize care delivery, understanding the sustainability and scalability of HaH models remain a critical area for ongoing research.

The rest of this paper is organized as follows. Section II describes the challenges and proposed solutions. The proposal is divided into a series of subproblems that are individually described and evaluated. Section III draws conclusions.

II. THE CHALLENGES AND PROPOSED SOLUTIONS

The goal of our hospital at home program is to produce the best outcomes at the lowest cost. Restated, the goal is to identify patients who are likely to succeed in the program without incurring significant resources, and to monitor those patients carefully and take action before any complications and negative outcomes. We have divided the hospital at home program into three distinct data problems. The first challenge is identifying and choosing the patients for the program. The second challenge is to predict the resources and cost needed to treat those patients and to predict the outcomes. This both helps us choose the best patients and provides information vital for resource planning and budgets. The third challenge is monitoring those patients during the program and intervening when necessary.

A. Choosing the Patient Cohort

For the purposes of our experiment, we have chosen to limit the program to patients who are primarily being treated by the cardiology specialty. Vast research has shown success in treating cardiology patients with hospital at home programs [6][7]. Furthermore, the technology to remotely monitor these patients is widely available and well researched [8][9]. To identify patients for the program, we leverage approaches from past research, and we follow the standards set by clinical trials for identifying and enrolling patients [10][11][12].

B. *Predicting the Outcomes and the Cost*

We set up a discrete event model for the hospital at home program. In such a model, the resources and constraints are identified. There is significantly less variation in available resources compared to creating a virtual digital twin of an entire hospital, but it is still a complex problem. Then, historical patient records are used to determine the resources needed to treat a patient. Finally, new patients are matched to similar past patient records using pattern matching and biostatistics. We propose to leverage the approach described in [13]. This will allow us to both predict outcomes (length of stay and readmissions primarily) and to estimate cost of treatment. Through modeling clinic needs of similar previous, we are able to show the expected length of stay, required therapies, medications and treatments, cost of treatment, and likelihood of adverse events or readmissions.

Finally, we leverage data from patients treated at home to create a supervised learning feedback loop. This will allow us to better refine the patient cohort selection program. To achieve this supervised learning, we propose to follow the approach of [14]. This allows us to data mining past information to best predict who will need intervention.

C. *Monitoring the Patients*

As already mentioned, there are significant resources and technology to monitor cardiac patients remotely. There are also many documented care paths tailed to cardiac patients [15][16]. We propose to leverage process mining tools to implement our care pathways and detect patient care variations [17][18].

Our remaining challenge becomes how to realize from all these data points when a patient is at risk and needs intervention. In [19], machine learning is used to predict adverse events in an ambulatory patient cohort. We propose that we retrain this specifically off of hospital-at-home patients and predict when a patient will need to return to acute hospital care. As the patient is more tightly monitored, and the number of adverse events is reduced, we believe that this approach has an opportunity for significant success.

III. CONCLUSION

Past results and research have demonstrated that hospital-at-home programs have the potential to improve outcomes and reduce costs. For this project, we look at how data and technology can help a health system develop a hospital at home program. We break this problem domain into several smaller and more controlled opportunities. We then propose to repurpose successful past research to approach each of the identified opportunities. We propose that these solutions can be applied to hospital-at-home programs and that there is significant opportunity to develop more successful programs using data and technology.

REFERENCES

[1] B. Leff et al., “Hospital at home: Feasibility and outcomes of a program to provide hospital-level care at home for acutely ill older patients,” Annals of Internal Medicine, pp. 27–35, 2022.

[2] A. Federman et al., “Association of a bundled hospital-at-home and 30-day postacute transitional care program with clinical outcomes and patient experiences,” in JAMA Internal Medicine, pp.1033–1041, 2018.

[3] D. Levine et al., ”Hospital-level care at home for acutely ill adults: A randomized controlled trial,” Annals of Internal Medicine, pp. 77–85, 2020.

[4] Centers for Medicare & Medicaid Services (CMS). (2020). “CMS announces comprehensive strategy to enhance hospital capacity amid COVID-19 surge,” (https://www.cms.gov).

[5] E. Hwang et al., “Hospital-at-home programs: Opportunities and challenges in emergency medicine” in Annals of Emergency Medicine, pp. 255–262, 2023.

[6] A. Cherukara et al., "Bringing heart care home: management of acute cardiovascular pathologies in the Home Hospital." European Heart Journal: Acute Cardiovascular Care 14.6 pp. 364-374, 2025.

[7] V. Tibaldi et al., "Hospital at home for elderly patients with acute decompensation of chronic heart failure: a prospective randomized controlled trial." Archives of internal medicine, pp. 1569-1575, 2009.

[8] A. Bui and G. Fonarow, "Home monitoring for heart failure management." Journal of the American College of Cardiology 59.2, pp. 97-104, 2012.

[9] A. Martínez et al., "A systematic review of the literature on home monitoring for patients with heart failure." Journal of telemedicine and telecare 12.5, pp. 234-241, 2006.

[10] L. Friedman et al., Fundamentals of clinical trials. springer, 2015.

[11] J. Zivin, "Understanding clinical trials." Scientific American 282.4, pp. 69-75, 2000.

[12] R. Miotto and C. Weng, "Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials." Journal of the American Medical Informatics Association 22.e1, pp. 141-150, 2015.

[13] J. McGlothlin et al., "Predicting Hospital Capacity and Efficiency." HEALTHINF, 2018.

[14] J. McGlothlin et al., "A Data Mining Tool and Process for Congenital Heart Defect Management." AMIA, 2018.

[15] E. Peterson et al., "Implementing critical pathways and a multidisciplinary team approach to cardiovascular disease management." The American journal of cardiology, 2008.

[16] N. Every et al., "Critical pathways: a review." Circulation 101.4, pp. 461-465, 2000.

[17] L. Perimal-Lewis, D. Lua, and C. Thompson, "Health intelligence: Discovering the process model using process mining by constructing Start-to-End patient journeys." Proceedings of the Seventh Australasian Workshop on Health Informatics and Knowledge Management-Volume 153, 2014.

[18] S. Behnam and O. Badreddin, "Toward a Care Process Metamodel: For business intelligence healthcare monitoring solutions." 2013 5th International Workshop on Software Engineering in Health Care (SEHC). IEEE, 2013.

[19] J. McGlothlin et al., "Predicting Adverse Events in Developmental Disabilities Population." HealthINF, 2025.

Leveraging Observational Medical Outcomes Partnership (OMOP) Data to Populate Disease Registries

James P. McGlothlin
RSM US LLP
Dallas, TX USA
jamie.mcglothlin@rsmus.com

Timothy Martens
The Heart Center
Cohen Children's Medical Center
Queens, NY
tmartens1@northwell.edu

Abstract— **Health care disease registries and procedural registries serve a vital purpose in support of research and patient quality. However, it requires a significant level of clinician effort to collect and submit the data required by each registry, and there are over 1000 common patient registries. In previous research, we have evaluated using supervised learning in conjunction with generative artificial intelligence to generate accurate content for disease registries. However, one of the largest challenges was to extract complete and meaningful data from the electronic medical record in a format that enabled the generative Artificial Intelligence (AI) tools. Standards like HL7 and Fast Healthcare Interoperability Resources (FHIR) were insufficient and burdensome. In this project, we propose using the new Observational Medical Outcomes Partnership (OMOP) data standard to acquire this data. Our Electronic Medical Record (EMR) software provides access to this data in the cloud without requiring extraction and transformation. The goal of this project is to utilize this data and technology to improve the population of disease registry records.**

Keywords- population health; OMOP; congenital heart disease; thoracic surgery; artificial intelligence.

I. INTRODUCTION

Patient registries are essential tools for improving healthcare quality, supporting clinical research, and ensuring patient safety. In the U.S. alone, there are over 1,000 active registries tracking patient outcomes and provider performance [1]. A mid-sized pediatric hospital in the U.S. identified 29 registries in which it actively participates, requiring more than 45,000 staff hours annually—including over 3,000 hours of physician time—for data abstraction. This reflects a significant investment of skilled clinical labor.

Despite the high resource demand, opting out of registry participation is not feasible. Registries are not only crucial for advancing research and public health, but they also influence financial incentives. Many registries contribute to provider and hospital performance ratings that impact reimbursement models, such as the Merit-Based Incentive Payment System (MIPS) from Centers for Medicare & Medicaid Services (CMS) [2].

Recent advances in generative AI and Large Language Models (LLMs) offer promising opportunities to reduce this burden [3]. Research has demonstrated that this technology can generate accurate structured information from unstructured clinical text without extensive retraining [4].

The greatest challenge has been to extract the complete patient record into clinical text useful for generative AI. Furthermore, the registry fields have to be explained in detail. In our previous research [3], we leveraged the FHIR interface to extract this data. However, we found that this required significant configuration and custom development, and did not provide the complete data view we needed to fully automate populating the disease registries.

In this proposal, we address this challenge by leveraging the OMOP. We access this data through a cloud interface with no extraction. We assert that this solution will improve the disease registry data and reduce the manual effort.

The rest of this paper is organized as follows. Section II describes the registry and problem domain. Section III describes the OMOP standard and how we apply this protocol. In Section IV, we present conclusions.

II. REGISTRY

For this research, we have decided to limit ourselves to one specific registry. The Society of Thoracic Surgeons (STS) Congenital Heart Surgery Database (CHSD) is a comprehensive, multicenter clinical registry developed to collect and analyze data on congenital heart operations in the United States and participating international institutions [11]. Established in 1994, the STS CHSD aims to improve the quality of care and outcomes for pediatric and adult patients with congenital heart disease by facilitating evidence-based practice through benchmarking, quality improvement, and clinical research. The database captures detailed procedural, demographic, and outcomes data on nearly all types of congenital heart surgeries, including perioperative morbidity, mortality, and resource utilization. The inclusion of both preoperative risk factors and postoperative outcomes supports accurate risk stratification and performance evaluation across participating centers.

The STS CHSD plays a pivotal role in advancing pediatric cardiac surgery by enabling collaborative research, supporting public reporting, and guiding institutional quality improvement efforts. It is one of the largest and most robust congenital heart surgery registries globally, with contributions from over 120 centers performing thousands of procedures annually. Risk-adjusted outcomes reporting is facilitated through the use of standardized nomenclature and

 ISBN: 978-1-68558-312-5

analytic models, such as the STAT (Society of Thoracic Surgeons-European Association for Cardio-Thoracic Surgery) Mortality Categories. The registry has informed numerous peer-reviewed publications, contributing to the development of clinical guidelines and performance standards. As a cornerstone of outcomes research in congenital cardiac surgery, the STS CHSD continues to evolve with data validation enhancements, increased interoperability with electronic medical records, and integration with other congenital heart disease registries to support lifelong patient care.

The STS CHSD registry is very complex. There are more than 1000 required fields. Recent audit data (from the 2022 update) indicates that across just 11 audited centers, approximately 9,128 individual data field entries were assessed for completeness and accuracy during a single harvest period [5].

III. OMOP

The OMOP Common Data Model (CDM) is a standardized data framework developed by the Observational Health Data Sciences and Informatics (OHDSI) initiative to facilitate the systematic analysis of disparate observational health data sources. OMOP enables the transformation of heterogeneous clinical data—such as EMRs, claims data, and disease registries—into a unified format with standardized terminologies and data structures. By harmonizing disparate datasets into the OMOP CDM, researchers and institutions can conduct large-scale, reproducible analyses across diverse populations and settings [6]. The model supports a wide range of research, from drug safety surveillance to comparative effectiveness studies, while leveraging standardized vocabularies like SNOMED CT [12], RxNorm [13], and Logical Observation Identifiers Names and Codes (LOINC) [14] to ensure semantic interoperability.

Utilizing OMOP for disease registries offers significant advantages in terms of scalability, interoperability, and analytical rigor. Disease-specific registries, such as those for cardiovascular disease, diabetes, or rare conditions, can be mapped to the OMOP CDM to facilitate multi-institutional research, enable longitudinal patient tracking, and integrate with broader health data networks. This transformation allows for the use of standardized analytic tools developed by the OHDSI community, including cohort definitions, prediction models, and outcome analysis frameworks, thereby accelerating hypothesis testing and real-world evidence generation. Aligning disease registries with OMOP facilitates regulatory-grade data analysis and supports initiatives such as learning health systems and precision medicine by creating interoperable data ecosystems capable of continuous quality improvement and discovery [7].

ATLAS is an open-source tool that facilitates the design and execution of analyses on standardized, patient-level, observational data in OMOP. “ATLAS enables users to define cohorts using concepts derived from standard vocabularies such as SNOMED, ICD, and LOINC, ensuring that all participants share a common understanding of clinical events and data formats.” [8].

Epic is the predominant EHR vendor in USA [15]. In the new version of Epic, patient records are materialized in OMOP format using the Microsoft Fabric cloud environment. Leveraging this solution allows us to access the data through shortcuts without creating additional API calls, extractions or transformations [9][10]. This enables us to apply generative AI tools directly to the OMOP data in order to generate the information to populate the registry.

IV. CONCLUSIONS

Historically, disease registries have been populated through manual chart abstraction. Recent advances in interoperability and generative AI have narrowed the gap towards automated record keeping but have still fallen short. In this project, we propose utilizing OMOP to complete the data flow and enable full automation.

REFERENCES

[1] R. Gliklich, N. Dreyer, and M. Leavy, eds. "Registries for evaluating patient outcomes: a user’s guide.", 2014.

[2] S. Blumenthal, "The Use of Clinical Registries in the United States: A Landscape Survey,” EGEMS, 2017, p. 26.

[3] J. McGlothlin and T. Martens, “Using Artificial Intelligence and Large Language Models to Reduce the Burden of Registry Participation,” in HealthINF, 2025.

[4] A. Thirunavukarasu et al., "Large language models in medicine." Nature medicine 29.8, 2023, pp. 1930-1940.

[5] J. Jacobs et al., "Introduction to the STS National Database Series: outcomes analysis, quality improvement, and patient safety," The Annals of thoracic surgery 100.6, 2015, pp. 1992-2000.

[6] I. Reinecke et al., "The usage of OHDSI OMOP–a scoping review," German Medical Data Sciences 2021: Digital Medicine: Recognize– Understand–Heal, 2021, pp. 95-103.

[7] P. Biedermann, "Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases." BMC medical research methodolog, 2021, p. 238.

[8] “Introduction to Atlas”, Albert Einstein College of Medicine, https://einsteinmed.edu/uploadedFiles/centers/ICTR/new/introduction-to-atlas-manual.pdf , 2021.

[9] D. Ghosh, Mastering Microsoft Fabric: SAASification of Analytics. Springer Nature, 2024.

[10] B. Mohapatra et al., "Data Integration, Data Export, and Analytics in Dataverse," Deep Dive into the Power Platform in the Age of Generative AI: Architectural Insights and Best Practices for Intelligent Business Solutions. Berkeley, CA: Apress, 2024, pp. 159-224.

[11] J. Jacobs, "The society of thoracic surgeons congenital heart surgery database: 2019 update on outcomes and quality." The Annals of thoracic surgery 107.3, 2019, pp. 691-704.

[12] E. Chang and J. Mostafa, "The use of SNOMED CT, 2013-2020: a literature review." Journal of the American Medical Informatics Association,2021, pp. 2017-2026.

[13] S. Liu et al., "RxNorm: prescription for electronic drug information exchange.", IT professional, 2005, pp. 17-23.

[14] C. McDonald, "LOINC, a universal standard for identifying laboratory observations: a 5-year update." Clinical chemistry, 2003, pp. 624-633.

[15] R. Johnson, "A comprehensive review of an electronic health record system soon to assume market ascendancy: EPIC." J Healthc Commun 1.4, 2016, p. 36.