German Medical Data Sciences 2025: GMDS Illuminates Health
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI251392

From Proprietary Printed Forms to Standardized Digital Exchange Formats – Care Transition Records to FHIR as an Example

Viktor WERLITZ[a,1], Lukas KLEYBOLTE[b], Sabahudin BALIC[a],
Elisabeth V. MESS[a,] Andreas MAHLER[c], Claudia REUTER[a],
and Alexandra TEYNOR[a]

[a] *Institute for agile Software Development, Technical University of Applied Sciences, Augsburg, Augsburg, Germany*
[b] *Applied Technologies of Language and Assistance Systems, Technical University of Applied Sciences Augsburg, Augsburg, Germany*
[c] *Digitization Staff Unit, University Hospital Augsburg, Augsburg, Germany*
ORCiD: VW 0009-0006-8965-6058, AT 0000-0002-8141-0428

Abstract. **Introduction**: The transition from proprietary, paper-based care transition records (CTRs) to standardized digital formats like HL7 FHIR remains a significant challenge for healthcare institutions. Variability in document layouts, coupled with slow adoption of new interoperability standards, complicates efforts to digitize patient records while preserving data integrity and privacy. **Methods**: This study presents a machine learning-based pipeline for automated information extraction from scanned CTRs. Synthetic training data was generated using a custom CTR generator. A Detectron2-based object detection model, integrated with LayoutParser for document structure analysis and Tesseract OCR for text recognition, was trained on this synthetic dataset. Checkbox detection was performed via an image-processing pipeline based on pixel density analysis. Extracted information was mapped to the FHIR-based PIO-ULB (Pflegeinformationsobjekt - Überleitungsbogen) format using a custom serialization tool. **Results**: A synthetic dataset of 10,000 CTR samples was used for training and evaluation. The model achieved high values for accuracy, precision, recall, and F1-score metrics for synthetic data (97%, 98%, 95%, 97%) and showed robust performance for real-world data (85%, 86%, 83%, 85%). Lower performance on real-world data was attributed to layout variability and scanning artifacts absent from the synthetic training set. **Discussion**: The results demonstrate the feasibility of using machine learning for automated extraction and standardization of CTRs, particularly when relying on synthetic data to overcome data privacy constraints during development. While accuracy declines with real-world document variability, the approach provides a possible interim solution for facilitating interoperability in healthcare documentation. Future work will focus on extending data generation to cover complex document layouts and integrating advanced OCR and handwriting recognition methods to further improve extraction performance.

Keywords. Machine Learning, ML, LayoutParser, care transition record, CTR, MIO, PIO-ULB, FHIR, HL7

[1] Corresponding Author: Viktor Werlitz, Viktor.Werlitz@tha.de

1. Introduction

1.1. Background

In recent years, the healthcare sector has seen a significant need to adopt new interoperability and data exchange standards, such as the Fast Healthcare Interoperability Resources (FHIR) [1]. In Germany specifically, standards such as Medical Information Objects (MIO) (e.g., vaccination sheets or maternity records) are being introduced, and efforts to implement them are being made [2]. These standards are designed to improve the flow of patient data between healthcare providers, enabling better care coordination, enhanced patient outcomes, and more efficient clinical workflows [3].

Nursing documents are also digitized in the form of MIOs. In this case the term care information objects (CIOs, German: Pflegeinformationsobjekte, short: PIOs) is used. The in December 2022 introduced FHIR standard for care transition records (CTR, German: Überleitungsbogen, short: ULB) aims to standardize the exchange format of CTRs [4]. CTRs typically contain critical information for transferring patients from one institution to another.

However, since the CIO-CTR is not mandatory, software manufacturers are reluctant to implement it, and its adoption lags. That means until the standard is legally binding, standardized and non-standardized formats (layout, content) will likely co-exist for a long time. This poses a challenge to those systems that will support the standard in the future as they need to deal with non-standard formats, which is difficult due to the variability of content and layout.

To address this challenge, this paper uses machine-learning techniques to investigate the automated conversion of legacy-format CTRs (scanned and machine-readable PDFs) into the CIO-CTR standard. The first page of CTRs in the University Hospital Augsburg (UHA) proprietary format is used as an example for this study. We aim to show which concrete challenges to face and present a solution together with a dataset for comparing our results and further research.

1.2. State of the Art

In the medical domain, converting unstructured or semi-structured texts into standardized formats has been a longstanding research objective [5]. In Germany, this topic became much more relevant in 2020 after the Hospital Future Act (KHZG) [6] was passed, which requires hospitals to be digitized as far as possible [7,8]. HL7 FHIR has established worldwide recognition as the international standard for digital documents in the medical sector [9]. In their paper, Bouh et al. [10] showed how challenging the process of digitizing paper documents into FHIR- conform digital standard documents is. They identified the extraction of accurate medical information from scanned documents as the most complex step, particularly when utilizing optical character recognition (OCR) to make the documents machine-readable.

Various AI-based methods have been proposed for medical text extraction. A comparison of popular approaches can be seen in Table 1.

Table 1. A comparison of recent approaches for solving information extraction and/or transformation from unstructured to structured formats regarding capabilities, strengths, and limitations.

Method	Capabilities	Strengths	Limitations
MedXN[11]	Extracts and normalizes medication names	High accuracy in medication extraction	Works only with medication data
MedTime[12]	Extracts and normalizes time-based patient treatment data	Recognizes and transforms time-related data	Works only for data during hospitalization
NLP2FHIR[13]	Pipeline that combines MedXN and MedTime	Recognizes time-related data and outputs FHIR	Quality depends on the chosen FHIR modules
Opiod2Fhir[14]	Transforms handwritten prescriptions into FHIR	Recognizes handwritten data and transforms outputs in FHIR	Works only with prescription forms
FHIR-GPT[15]	Transforms clinical texts to FHIR medication statements	No fine-tuning is nee,ded and better results than NLP2FHIR	Works only with medications

These approaches focus on specific patient information or contexts. Initial literature research on the topic found that there are barely any approaches for use cases specific to the care domain such as the care transition records, hence a more specified and tailored approach for this project is needed.

1.3. Challenges in Test Data Acquisition – A Possible GDPR Conform Approach

Developing AI-based systems requires high-quality training data, which is challenging to obtain in the medical or care domain due to strict data protection and privacy regulations in Germany. One way to approach this problem is by creating synthetic dataSynthetic data refers to information that is artificially generated and is typically created using algorithms to replicate actual data. In terms of privacy and data scarcity this approach offers a solution because it is possible to mirror real patient information without compromising individual privacy. Additionally, researchers can develop and test AI-based systems without accessing sensitive personal data [16].

In the context of this work, a tool was implemented, which generates data safety regulation-compliant CTR forms in the original format of the UHA using anonymized patient data. These generated forms could then be used to train a machine learning (ML) model for layout parsing and optical character recognition (OCR), which enables automated extraction of information from CTRs. Finally, this extracted information can be mapped to the fields of the CIO-CTR. This new tool will be referred to as "CTR generator" and checking its results and further possible improvements of this tool for better performance of the ML model is part of this paper. Our methodological approach is shown in the following.

2. Methods

2.1. Synthetic CTR Generator for GDPR Compliant Training Data

CTRs consist mainly of person-related data and additional relevant patient information, such as vital parameters, allergies, involved health care practitioners, and contact persons.

To simulate scanned PDFs, the CTR Generator first creates HTML files with CTR layout, which are then converted into PNG format. For each CTR image produced by the data generator, a corresponding COCO (Common Objects in Context) [17] file is generated that contains both basic information about the image (width, height, and filename) and the bounding box label information from the HTML file for specific content areas.

The HTML file consists of multiple HTML sections, each containing HTML and CSS data that mimics the original UHA-CTR and bounding box information for corresponding HTML elements. To fill out the HTML sections with content, the CTR-Generator imports 200 rows of anonymized patient data from a CSV file provided by UHA and then populates the HTML template with text. As person-related data such as names and insurance numbers cannot be used, they are generated synthetically using Python-Faker. This synthetic data is embedded into the HTML template, ensuring compliance with the Patient Data Protection Act.

The generator is configurable to allow users to specify the number of synthetic UHA-CTRs to generate, along with a desired text length for free-text fields. This flexibility ensures that the generated dataset contains UHA-CTRs of varying lengths, preparing the classifier to handle diverse real-world scenarios effectively. The current generator setup ensures that the training process is aligned with the maximum section and element sizes found in actual UHA-CTRs, mimicking real-world data.

2.2. Layout Parsing, OCR, and checkbox recognition

The structure of CTRs is usually semi-structured and comprises various sections with different layout elements that contain specific information and may interact with each other. However, traditional OCR processes often overlook this structural information, extracting only textual content without preserving the relationships between layout components [18]. For instance, tables are reduced to individual cell text, losing their structural context. To address this limitation, we utilized the Python library LayoutParser [19] a toolkit that detects and organizes layout elements based on their positional relationships. By defining areas of interest, such as sections and table cells, and associating them through spatial heuristics, LayoutParser helps preserve structural context. For example, in a form containing a label and a text field, LayoutParser detects each as a separate bounding box. The label's position relative to the text field can be used to infer a parent-child relationship, preserving document structure. Each detected object retains the extracted text from its corresponding area, maintaining the original layout's context. LayoutParser also offers automatic layout recognition, which works with the vision library from Meta, Detectron2 [19]. We trained a model using the labeled CTRs generated by the CTR-Generator to automatically identify layout elements along with their corresponding labeled classes and relationships. This approach enables us to effectively address the challenge of variable layout sections in different sizes depending on the textual content of the layout blocks.

For extracting the textual information in the CTRs we used Google's OCR engine Tesseract [20], which enables us to get the content of text fields and table cells e.g..

Checkboxes in UHA-CTRs are commonly used to encode binary information but pose challenges for OCR due to their graphical nature [21]. To solve this, we applied an image preprocessing pipeline: first, the contrast and sharpness were enhanced, followed by binary thresholding. The pixel density within the checkbox bounding box was then analyzed—higher density indicated a checked box, while lower density indicated an unchecked box.

2.3. Standardization: Mapping Extracted Data to CIO

The next step in the workflow is transforming the extracted data into the CIO-CTR format. The values identified by the layout parsing and checkbox recognition module must be mapped to CIO-CTR conform data structures (resources). These data structures are then serialized to standard CIO-CTR XML files. This serialisation is done by using a software solution [22], which we developed in previous work [23].

Another key functionality of the utilized software lies in its ability to load the extracted data into relevant fields of a CIO-CTR template and visualize it. The tool's user interface allows users to review and modify the extracted data, ensuring that any potential errors or discrepancies can be corrected before the final CIO is saved.

3. Results

The synthetic dataset (N=10,000, available upon request) was split into 60% training, 20% validation, and 20% test data. Each CTR sample consists of an annotated PNG image with structured metadata stored in a COCO JSON file. The JSON file includes bounding box coordinates, label entities, and hierarchical relationships between form elements. For testing the performance of the system on real-world data, a set of CTRs (n = 10) was created manually by the UHA using their in-house clinical information system using realistic but made-up patient data. This data was used on the model and manually evaluated afterwards using a tool developed to display FHIR-based CIO data [22].

To evaluate the performance of the Detectron2-based model in recognizing and extracting data from CTRs, a field is considered correctly extracted only if both its text content and positional accuracy match the ground truth, while partial matches (e.g., minor OCR errors or misplaced bounding boxes) are counted as incorrect. This strict evaluation is necessary to ensure the model only extracts correct content without any changes and discards any unreliable extractions. To interpret the results based on this evaluation, we computed several standard metrics used in object detection and text recognition tasks. These results can be seen in Table 2.

Table 2. A comparison of the model's performance on real-world and synthetic data.

Metric	Synthetic data	Real-world data
Accuracy	97%	85%
Precision.	98%	86%
Recall	95%	83%
F1-Score	97%	85%

The results show that using the model to analyze the real-world CTRs results in lower performance, as these CTRs had some graphical anomalies not present in the synthetic training data (e.g. information distributed over page breaks). Since this is a realistic issue, the synthetic data generation should be extended to also cover this issue. The metrics indicate that the model is a promising approach in detecting relevant data fields within the synthetic training dataset and performs well in distinguishing between different information categories. The high precision (86%) indicates that the model minimizes false positives, successfully distinguishing between relevant and irrelevant fields. However, the recall (83%) suggests that some fields were missed, likely due to variations in layout structure or OCR misinterpretation. The F1-score (84.5%) balances both metrics, confirming the overall effectiveness of the approach.

4. Discussion

The transition from legacy formats (e.g., PDFs, paper-based records) to structured digital standards such as FHIR and MIO is essential for improving interoperability and patient care efficiency. Our study demonstrates that machine learning-based extraction, particularly using Detectron2 for layout analysis and OCR, can serve as a viable bridge to facilitate this process. By leveraging synthetically generated training data, we mitigate the challenges associated with data privacy while ensuring the model learns to recognize document layouts specific to the first page of CTRs of UHA. This prototype serves as a proof-of-concept, highlighting that automated extraction is feasible, accelerating the adoption of interoperability standards in real-world applications. This approach is especially effective as an add-on tool for institutions handling incoming CTRs from other facilities since parts of the data can be automatically extracted before being finalized and adopted by the care workers as was tried and tested in another study including regional care facilities [23].

Despite the promising results, several factors complicate the extraction of structured information from CTRs:

- **Inconsistent Document Layouts**: Variations in formatting, spacing, and alignment across different CTR instances introduce noise into the detection process. Traditional OCR methods struggle with text appearing in irregular positions, necessitating robust layout parsing [24]. Preprocessing techniques such as denoising filters or inpainting models could potentially recover partially lost text.
- **Degraded Scans and Cut-Off Sections**: Scanned documents often suffer from artifacts such as blurring, ink smudging, and cropped content due to scanning inconsistencies. This results in partial loss of information, requiring additional post-processing steps to infer missing data reliably.
- **Checkbox Recognition**: Recognizing checkboxes presents a unique challenge, as they can vary in size, shape, and position within the document. Traditional OCR does not always accurately distinguish between selected and unselected checkboxes, making classification less reliable without additional heuristics [21].
- **Handwritten Annotations**: Some CTRs include handwritten notes or signatures, which standard OCR models struggle to interpret accurately. Future iterations of the system could integrate handwriting recognition models to enhance extraction capabilities.

While some aspects of automatic recognition remain challenging, several elements of CTRs are well-structured, making them easier to process:

- **Consistent Header and Patient Information Fields**: The primary patient information, including name, birthdate, and institutional details, follows a relatively fixed layout, allowing high accuracy in text extraction.

- **Fixed-Position Tables**: Certain sections, such as medication lists and practitioner details, are presented in structured tables, making them easier to detect and extract accurately using layout parsing techniques.
- **Standardized Institutional Data**: Sender and recipient institutions follow predefined formats, enabling rule-based validation to improve extraction accuracy further and reduce errors.

These easy-to-recognize elements make the approach worthwhile, as they ensure that a substantial portion of critical patient data can be reliably extracted and mapped to CIOs, even when other aspects of the document pose recognition difficulties. Future work will focus on refining the model to improve recognition accuracy by inserting data with artifacts or cropped content and integrating complementary techniques such as transformer-based OCR for contextual extraction. Additionally, the confidence of the model should be mentioned - a value ranging from 0 to 1 describing how certain the model is about its prediction. This metric should be used in future work to analyze wrong predictions for further insights. Finally, extending the system to support the remaining pages of UHA CTRs and CTRs of other institutions, in general, will be crucial for achieving a more comprehensive transformation into the CIO-CTR standard.

5. Conclusion

The digital transformation of healthcare is imperative, yet many institutions still rely on paper-based records for now. The adoption of structured standards, such as FHIR and MIO, is essential for interoperability, but the transition is slow and complex. Since full adoption of digital standards will take time, interim solutions that facilitate integration between structured and unstructured formats are crucial. Machine learning-based approaches can act as a bridge, enabling institutions to transition without immediate full-scale adoption.

Our study demonstrates that layout parsing using Detectron2 is effective for extracting structured data from CTRs, achieving high precision in structured field recognition. However, model performance declines as document diversity increases, particularly in handling variable layouts, degraded scans, and checkbox detection. In healthcare applications, reliability and accuracy are paramount. The feasibility study presented in this paper highlights that while automated recognition is a promising solution for standardizing unstructured data, further refinement and optimization are required. Future research should focus on enhancing model robustness against layout variability, degraded document quality, and non-text elements (e.g., checkboxes and handwritten annotations). Expanding support for additional healthcare document types will be critical for achieving widespread adoption in clinical settings.

Declarations

Conflict of Interest: The authors declare no conflict of interest.
Author contributions: VW, LK, SB, MR, AT and AM were involved in the concept of work, data acquisition, and interpretation. VW, LK and SB were involved in the development of the software. VW, LK, SB and AT were involved in writing the

manuscript. VW, EVM and AT were involved in substantial revising of the manuscript. All authors approved the manuscript in the submitted version and take responsibility for the scientific integrity of the work.

Acknowledgment: This research is part of the project CARE REGIO, funded by the Bavarian State Ministry of Health, Care and Prevention. We want to thank our collaborating facilities for their time and effort.

References

[1] C.N. Vorisek, M. Lehne, S.A.I. Klopfenstein, P.J. Mayer, A. Bartschke, T. Haese, and S. Thun, Fast Healthcare Interoperability Resources (FHIR) for Interoperability in Health Research: Systematic Review. JMIR Med Inform 10 (2022), e35724.

[2] KBV, Medizinische Informationsobjekte, https://www.kbv.de/html/mio.php.

[3] eCQI, FHIR® - Fast Healthcare Interoperability Resources®, https://ecqi.healthit.gov/fhir [cited 2025 February 21].

[4] Kassenärztlichen Bundesvereinigung, ÜBERLEITUNGSBOGEN 1.0.0, https://mio.kbv.de/pages/viewpage.action?pageId=73138833 [cited 2025 February 24].

[5] H.-J. Kong, Managing Unstructured Big Data in Healthcare System. Healthc Inform Res 25 (2019), 1–2.

[6] *Krankenhauszukunftsgesetz – KHZG.*

[7] J. Pavão, R. Bastardo, and N.P. Rocha, A systematic review of the use of FHIR to support clinical research, public health and medical education. DTA ahead-of-print (2024).

[8] J. Sass, S. Zabka, A. Essenwanger, J. Schepers, M. Boeker, and S. Thun, Fast Healthcare Interoperability Resources (FHIR®) Representation of Medication Data Derived from German Procedure Classification Codes (OPS) Using Identification of Medicinal Products (IDMP) Compliant Terminology. Stud Health Technol Inform 278 (2021), 231–236.

[9] Assistant Secretary for Technology Policy, Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR), https://www.healthit.gov/topic/standards-technology/standards/fhir.

[10] M.M. Bouh, F. Hossain, and A. Ahmed, A Machine Learning Approach to Digitize Medical History and Archive in a Standard Format, pp. 230–236.

[11] S. Sohn, C. Clark, S.R. Halgrim, S.P. Murphy, C.G. Chute, and H. Liu, MedXN: an open source medication extraction and normalization tool for clinical text. J Am Med Inform Assoc 21 (2014), 858–865.

[12] Y.-K. Lin, H. Chen, and R.A. Brown, MedTime: a temporal information extraction system for clinical narratives. J Biomed Inform 46 Suppl (2013), S20-S28.

[13] N. Hong, A. Wen, F. Shen, S. Sohn, S. Liu, H. Liu, and G. Jiang, Integrating Structured and Unstructured EHR Data Using an FHIR-based Type System: A Case Study with Medication Data. AMIA Jt Summits Transl Sci Proc 2017 (2018), 74–83.

[14] J. Wang, W.C. Mathews, H.A. Pham, H. Xu, and Y. Zhang, Opioid2FHIR: A system for extracting FHIR-compatible opioid prescriptions from clinical text, pp. 1748–1751.

[15] Y. Li, H. Wang, H.Z. Yerebakan, Y. Shinagawa, and Y. Luo, *FHIR-GPT Enhances Health Interoperability with Large Language Models,* 2023.

[16] J.M. Abowd and J. Lane, New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Centers, pp. 282–289.

[17] COCO Consortium, Data format, https://cocodataset.org/#format-data [cited 2024 November 12].

[18] S. Pallavi, R.R. Pranesh, and S. Kumar, *A Conglomerate of Multiple OCR Table Detection and Extraction,* arXiv, 2020.

[19] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick, *Detectron2,* 2019.

[20] R. Smith, An Overview of the Tesseract OCR Engine, pp. 629–633.

[21] J.M. Istle, *Optical character recognition for checkbox detection,* 2004.

[22] THA_IAS, THA_IAS Github Projects, https://github.com/THAias.

[23] M. Regner, V. Werlitz, L. Kleybolte, E.V. Mess, S. Balic, L. Daufratshofer, S. Tilmes, A. Mahler, P. Heidegger, C. Reuter, and A. Teynor, Entwicklung eines Editors zur Erstellung und Bearbeitung Pflegerischer Informationsobjekte (PIOs) zur Pflegeüberleitung. GMS Medizinische Informatik, Biometrie und Epidemiologie 20 (2024).

[24] A. Fateh, M. Fateh, and V. Abolghasemi, Enhancing optical character recognition: Efficient techniques for document layout analysis and text line detection. Engineering Reports 6 (2024).