

Article

Text2FHIRwallet: Automated Generation of FHIR Patient Summaries from Unstructured Cardiology Reports Using Fine-Tuned Portuguese Language Models—Development and Evaluation of a Health Professional Wallet

João C. Ferreira ^{1,2,3,*} , Isabel Rosa ⁴ and Ricardo Correia ⁵ 

¹ Department of Information Science and Technology, ISTAR, Instituto Universitário de Lisboa (ISCTE-IUL), Avenida das Forças Armadas, 1649-026 Lisbon, Portugal

² Faculty of Logistics, Molde University College, Britvegen 4, NO-6410 Molde, Norway

³ INESC INOV—Instituto de Novas Tecnologias, Rua Alves Redol 9, 1000-029 Lisbon, Portugal

⁴ Genesis Digital Solutions, Rua Soeiro Pereira Gomes, Lote 1, 7B, 1600-198 Lisbon, Portugal

⁵ BioGHP, Lisbon, Portugal

* Correspondence: joam@himolde.no

Abstract

Background and Objectives: Cardiology departments generate large volumes of unstructured free-text reports that impose substantial manual review burdens on clinicians; at Hospital de Santa Maria—Portugal’s largest public hospital—manual review of 12,651 reports took approximately seven minutes per report, representing over 1475 h of avoidable administrative work. This study presents Text2FHIRwallet, a health professional digital wallet that automates extraction and structuring of clinical entities from unstructured Portuguese cardiology reports using fine-tuned Named Entity Recognition (NER) models and maps the results to Fast Healthcare Interoperability Resources (FHIR) R4 patient summaries. **Materials and Methods:** Following the Design Science Research Methodology (DSRM) and CRISP-DM, we fine-tuned four transformer-based models—BERTimbau Base, BERTimbau Large, Albertina PT-PT, and MediAlbertina—on 305 manually annotated cardiology reports (77,309 tokens; $\kappa = 0.85$ inter-annotator agreement) covering eight clinical entity types, drawn from a corpus of 12,651 anonymised documents. Entities were mapped to FHIR R4 resources and delivered through a secure, role-based mobile wallet (React Native). Evaluation comprised token-level NER benchmarking with bootstrapped confidence intervals and McNemar’s testing, FHIR mapping accuracy assessment on 100 manually reviewed reports, processing-efficiency measurement, and a usability pilot with 10 cardiologists (SUS, NPS). **Results:** MediAlbertina achieved the highest NER performance (macro F1 = 0.985, 95% CI: 0.979–0.990), significantly outperforming all baseline models ($p < 0.01$, McNemar’s test) and comparing favourably with—though not directly comparable to, given differing languages and datasets—published benchmarks such as GPT-4 (F1 = 0.962 in ophthalmology NER) and fine-tuned BERT models for lung cancer NER (F1 $\approx$ 0.85–0.90). FHIR mapping accuracy was 98% on 100 independently reviewed reports. Report processing time was reduced from approximately seven minutes to 15–30 s (93–96% reduction), with peak batch-inference throughput of up to 1000 reports/h under parallelised GPU load (observed end-to-end throughput in pilot deployment was approximately 250 reports/h). The pilot usability evaluation yielded a SUS score of 87 (excellent) and an NPS of 80. **Conclusions:** Text2FHIRwallet demonstrates that domain-specific fine-tuning of a Portuguese-language pretrained language model achieves near-ceiling clinical NER accuracy, enabling scalable, interoperable, and privacy-compliant patient summary generation from unstructured cardiology text, offering an end-to-end pathway for integrating



Academic Editor: Christos Bouras

Received: 4 July 2026

Revised: 24 July 2026

Accepted: 3 August 2026

Published: 5 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

AI-driven NLP into clinical workflows and FHIR-based health information ecosystems, with implications for administrative efficiency, care coordination, and clinical research in non-English-language settings.

Keywords: pre-trained language models; named entity recognition; FHIR; patient summaries; cardiology; digital health wallet; natural language processing; clinical NLP; Portuguese language models

1. Introduction

The integration of Artificial Intelligence (AI)—and large language models (LLMs) in particular—into clinical practice is reshaping how healthcare systems manage and extract value from unstructured medical data [1,2]. In cardiology, clinicians face an exponential growth in electronic health record (EHR) documentation, imaging reports, and encounter notes that must be reviewed, synthesised, and acted upon within constrained timeframes [3,4]. This documentation burden directly competes with time available for patient care and translates into measurable institutional costs: increased administrative workload, delayed information retrieval, and barriers to systematic data reuse for research and quality improvement [5,6].

Natural Language Processing (NLP), powered by transformer-based LLMs, offers a technically mature solution for converting unstructured clinical text into structured, queryable representations [7,8]. Clinical Named Entity Recognition (NER)—the automated identification of clinical concepts such as diagnoses, medications, symptoms, and vital signs from free-text reports—has demonstrated strong performance across multiple specialties [9,10]. Concurrently, the Fast Healthcare Interoperability Resources (FHIR) standard has emerged as the dominant framework for structured health data exchange, enabling clinical information to flow across EHR systems and support downstream analytics, decision support, and cohort research [11,12].

This research was motivated by a direct clinical partnership with cardiologists at Hospital de Santa Maria, Lisbon's largest public hospital. The clinical team identified a persistent bottleneck: manual review of 12,651 anonymised Portuguese-language cardiology reports—collected between November 2020 and December 2022 under ethical approval (reference 132/22)—required approximately seven minutes per document. Extrapolated across the full corpus, this represented over 1475 h of clinician time devoted to unstructured text review rather than patient care. The absence of structured, interoperable patient summaries further impeded cohort studies, disease progression tracking, and multi-professional coordination.

In response, we developed Text2FHIRwallet: a health professional digital wallet that automates the end-to-end transformation of unstructured Portuguese cardiology reports into FHIR R4-compliant patient summaries using fine-tuned transformer-based NER models, delivering structured data through a secure mobile application. The system was designed to meet three core clinical requirements: (i) automated extraction of clinical entities from unstructured Portuguese text; (ii) generation of interoperable FHIR patient summaries aligned with the International Patient Summary (IPS) profile; and (iii) secure, professional-grade access, sharing, and audit of those summaries via a digital wallet interface.

The existing literature demonstrates strong NER performance by LLMs in English-language clinical contexts [9,12,13], but Portuguese-language applications—particularly in cardiology—remain substantially underexplored. Prior FHIR integration work has addressed medication extraction [11] and phenotyping [14] but has not produced validated

end-to-end pipelines combining LLM-based NER, FHIR mapping, and clinician-facing wallet delivery. Digital health wallets have been proposed for patient-centric data management [15], but professional-oriented AI-driven wallet systems for clinical report structuring have not been previously reported. Text2FHIRwallet directly addresses this intersection.

This paper makes the following contributions:

- Design and evaluation of a complete end-to-end pipeline from unstructured Portuguese cardiology text to FHIR R4-compliant patient summaries.
- Fine-tuning and benchmarking of MediAlbertina—a domain-adapted Portuguese pretrained language model—achieving state-of-the-art cardiology NER performance (macro F1 = 0.985) with statistically validated superiority over all baselines.
- Development and usability evaluation of a secure health professional wallet application achieving excellent clinician acceptance (SUS = 87, NPS = 80).
- Empirical evaluation of the system across routine care, emergency triage, and research cohort analysis use cases at a single tertiary cardiology centre.

The remainder of this paper is structured as follows: Section 2 presents the systematic literature review. Section 3 describes the methodology and system architecture. Section 4 presents demonstration and evaluation results. Section 5 provides discussion, including limitations and ethical considerations. Section 6 concludes the paper.

2. Literature Review

This literature review follows the PRISMA 2020 guidelines [16] to ensure transparency and reproducibility in study selection and synthesis.

2.1. Search Strategy and Selection

A systematic search was conducted on PubMed, Scopus, IEEE Xplore, and ACM Digital Library using the following query strings: “large language models named entity recognition medical reports”; “fine-tuning LLMs for structuring cardiology reports”; “FHIR integration with natural language processing in healthcare”; and “digital health wallet verifiable credentials”. The search covered publications from 2019 to 2025. After duplicate removal, 39 records were screened based on titles and abstracts. Twenty-five full-text articles were assessed for eligibility; ten studies were ultimately included in the qualitative synthesis based on direct relevance to LLMs in clinical NER, FHIR integration, or health data wallets. Exclusions were due to: absence of AI/LLM focus ($n = 10$), non-medical domains ($n = 3$), or insufficient methodological detail ($n = 2$).

2.2. LLMs for Named Entity Recognition in Clinical Texts

The literature increasingly demonstrates the utility of LLMs for NER in unstructured medical reports, though performance varies substantially with model architecture, training strategy, and language. Hu et al. [9] evaluated GPT-3.5 and GPT-4 for clinical NER, achieving relaxed F1-scores of 0.726 and 0.861, respectively; prompt engineering improved GPT-4 performance by 20%, yet results remained below fine-tuned BERT baselines on structured extraction tasks, suggesting that hybrid approaches combining prompt engineering with domain fine-tuning may be optimal.

Majid et al. [13] compared GPT-4 (F1 = 0.962) with BERT variants including BioBERT (F1 = 0.875) for medication NER in ophthalmology notes, demonstrating GPT-4 superiority in domain-specific extraction tasks under zero-shot conditions. Arzideh et al. [17] compared encoder-only models with generative LLMs for NER in lung cancer reports and found that fine-tuned BERT variants (F1 $\approx$ 0.85–0.90) consistently outperformed zero-shot generative approaches, underscoring the value of domain-specific fine-tuning for structured entity extraction. Šuvalov et al. [18] demonstrated that synthetic data generation via LLMs

can significantly expand limited training corpora, improving NER performance in low-resource clinical settings—a finding directly relevant to minority-language corpora such as European Portuguese. Ogi et al. [19] explored prompt-based LLM approaches for Japanese medical incident reports, with GPT-4 achieving competitive performance and suggesting cross-lingual generalisability.

2.3. FHIR Integration with NLP and LLMs

FHIR has emerged as the dominant framework for structured health data exchange, and several recent works have demonstrated the feasibility of NLP-driven FHIR resource generation. Li et al. [11] developed FHIR-GPT, achieving greater-than-90% exact match rates for medication statement extraction and improving NLP pipeline performance by 3–50% across dosage and timing elements. Hong et al. [14] demonstrated an NLP2FHIR phenotyping framework that generated 244,438 FHIR instances from discharge summaries with F1-scores up to 0.9536 for obesity classification. Wen et al. [20] introduced CQL4NLP, integrating FHIR NLP extensions into Clinical Quality Language for EHR-driven phenotyping, demonstrating extensibility for AI-augmented clinical workflows. Hou et al. [21] proposed LINK-FHIR, an LLM-integrated normalisation framework improving entity resolution in complex clinical text. Collectively, these works establish FHIR as a robust bridge for AI-structured clinical data, though none addresses Portuguese-language cardiology or integrates an end-to-end professional wallet delivery layer.

2.4. Digital Health Wallets for Health Professionals

Professional-oriented digital health wallets remain an emerging domain. Harrell et al. [15] designed MediLinker, a blockchain-based wallet for verifiable credential storage supporting secure health data sharing and professional permission management. Elvas et al. [22] explored health data sharing systems enabling knowledge creation across clinical and research settings, demonstrating the value of structured data exchange for longitudinal analytics. While these works confirm demand for professional-grade wallet systems, neither incorporates AI-driven report structuring—a gap that Text2FHIRwallet addresses.

2.5. Synthesis and Research Gap

The reviewed literature demonstrates strong evidence for LLMs in clinical NER and for FHIR as an interoperability standard, but reveals three persistent gaps: (i) limited cardiology-specific NER work in non-English languages, particularly Portuguese; (ii) absence of validated end-to-end pipelines integrating LLM-based NER, FHIR mapping, and clinical wallet delivery; and (iii) no AI-enabled professional wallet systems for clinical report structuring. Text2FHIRwallet is designed to address all three gaps through a validated, end-to-end system evaluated at a tertiary cardiology centre.

3. Materials and Methods

This study adopts the Design Science Research Methodology (DSRM) as proposed by Peffers et al. [23], combined with the Cross-Industry Standard Process for Data Mining (CRISP-DM) for model development. DSRM provides a structured framework for designing, developing, and evaluating information technology artefacts in applied research contexts. This section covers DSRM Phases 1–3 (problem identification, objective definition, and design and development); DSRM Phases 4–5 (demonstration and evaluation) are reported in Section 4.

3.1. Problem Identification and Motivation (DSRM Phase 1)

The clinical motivation for Text2FHIRwallet was established through direct stakeholder engagement with cardiologists at Hospital de Santa Maria. The dataset under study comprised 12,651 anonymised cardiology reports in European Portuguese (RTF format, Latin-1 encoding), collected between 27 November 2020 and 31 December 2022 under ethical approval (reference 132/22, Ethics Committee of Hospital de Santa Maria). Manual review of each report required approximately seven minutes per document—representing over 1475 h of clinician time across the full corpus. The absence of structured, interoperable patient summaries additionally impeded cohort research, disease progression tracking, and multi-professional coordination. Iterative consultations with clinical staff confirmed the need for an automated structuring solution preserving clinical accuracy, ensuring GDPR-compliant data privacy, and integrating with FHIR-based hospital infrastructure.

3.2. System Requirements (DSRM Phase 2)

System requirements were elicited through iterative consultations with cardiologists, analysis of the 12,651-report corpus, and review of the related literature. Functional requirements are summarised in Table 1, non-functional requirements in Table 2, and the DSRM objective mapping in Table 3.

Table 1. Core functional capabilities of Text2FHIRwallet.

Capability	Description
Text Ingestion and Pre-Processing	Automated ingestion of unstructured Portuguese cardiology reports (RTF/PDF) via REST API; Latin-1-to-UTF-8 conversion; text normalisation for tokenisation.
Pretrained Language Model-Based NER	Fine-tuned transformer models identify eight clinical entity types: Disease, Symptom, Medication, Dosage, Vital Sign, Medical Procedure, Outcome, and Progress.
FHIR Mapping and Summary Generation	Extracted entities mapped to FHIR R4 resources (Patient, Condition, Observation, MedicationStatement) assembled into Bundles conforming to the International Patient Summary (IPS) profile.
Secure Storage and Retrieval	Encrypted PostgreSQL database with HAPI FHIR extensions; wallet interface supporting structured search by patient identifier, clinical keyword, or temporal trend.
Collaboration and Access Control	Role-based access control for secure summary sharing among clinicians, with immutable audit logging.
Continuous Learning	Clinician feedback loops enabling correction of NER errors and scheduled model retraining for ongoing quality improvement.

Table 2. Non-functional requirements and performance targets.

Dimension	Target/Specification
Accuracy	NER macro F1 ≥ 0.95 ; FHIR mapping accuracy $\geq 95\%$
Processing Efficiency	Per-report latency < 60 s; throughput ≥ 1000 reports/h on GPU infrastructure
Security and Privacy	AES-256 encryption at rest and in transit; GDPR compliance; tamper-proof blockchain-inspired audit trail
Interoperability	Full conformance to FHIR R4 and IPS profile for EHR integration
Usability	SUS ≥ 80 (good); pilot evaluation with ≥ 10 clinicians

Table 3. System requirements mapped to DSRM objectives.

Category	Representative Requirement	DSRM Objective
Functional	NER extraction with fine-tuned pretrained language model	Automate report structuring to reduce manual workload
Non-Functional	FHIR R4 compliance; AES-256 encryption	Ensure interoperability and data security
User Experience	Mobile wallet with structured search and feedback mechanism	Enhance usability for health professionals

3.3. System Architecture (DSRM Phase 3)

3.3.1. High-Level Architecture

Text2FHIRwallet employs a modular, microservice-based architecture comprising four layers: (1) Data Ingestion, (2) AI Processing, (3) Storage and Management, and (4) User Interface. Data flows from unstructured cardiology reports through a fine-tuned NER pipeline to FHIR-compliant resources, stored securely and accessed via a mobile wallet. The architecture is illustrated in Figure 1.

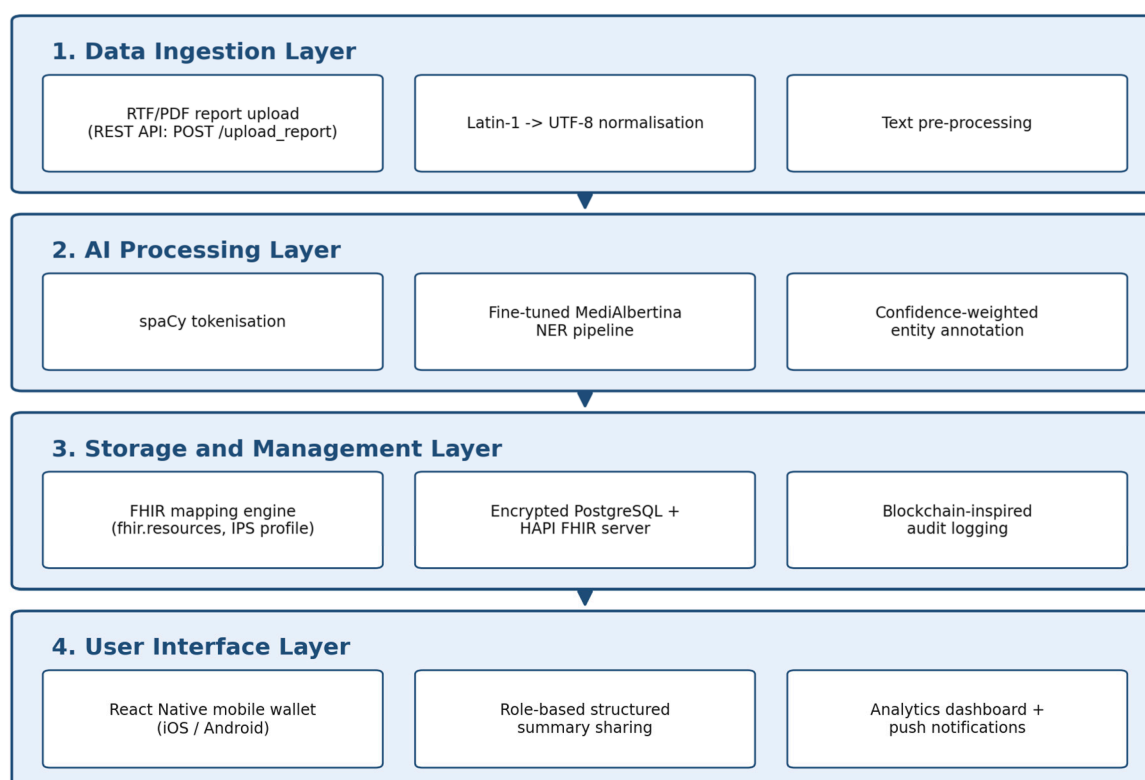


Figure 1. Text2FHIRwallet four-layer system architecture. Data flows from unstructured cardiology report ingestion (Layer 1) through the fine-tuned NER pipeline (Layer 2) to FHIR R4 resource generation, encrypted persistence, and audit logging (Layer 3), and is delivered to health professionals through the mobile wallet interface (Layer 4).

Data Ingestion Layer: This layer manages report upload or synchronisation from hospital systems, including RTF-to-text conversion and Latin-1 encoding normalisation. Reports are submitted via a REST API endpoint (POST/upload_report).

AI Processing Layer: This layer executes the fine-tuned NER pipeline. Reports are tokenised using spaCy and processed by the fine-tuned MediAlbertina model. During local prototyping, this was served via the Hugging Face Inference API; the intended production configuration is the self-hosted GPU infrastructure described in Sections 3.3.3 and 5.5, so that clinical text does not leave institutionally controlled infrastructure at inference

time—this is pending final confirmation against the current system build (see Section 5.5 for the current status of that confirmation). Confidence-weighted entity annotations are passed to the FHIR mapping engine.

Storage and Management Layer: An encrypted PostgreSQL database with HAPI FHIR server extensions provides FHIR resource persistence. A Hyperledger Fabric-inspired blockchain architecture provides tamper-proof access event logging, with no personally identifiable data stored on-chain.

User Interface Layer: A React Native mobile application (iOS and Android) delivers the wallet interface, including structured patient summary search, selective sharing with role-based access control, an analytics dashboard, push notifications for high-priority alerts, and offline caching for emergency use. Access is secured via JWT authentication.

3.3.2. Key Components

NER Module: MediAlbertina was selected as the primary model based on evaluation results. Fine-tuning was performed using the Hugging Face Transformers library with the following hyperparameters: learning rate = 2×10^{-5} ; batch size = 16; and epochs = 3, with early stopping on validation F1. Input tokenisation used the spaCy tokeniser; annotations were produced using Prodigy by two independent clinical annotators.

FHIR Mapping Engine: Custom Python scripts map NER entities to FHIR R4 resources—for example, Disease → Condition; Symptom → Observation; Medication → MedicationStatement; and Dosage → Observation extension. Schema validation is performed using the fhir.resources Python library against IPS profile constraints.

Vital Sign, Medical Procedure, Outcome, and Progress entities are likewise mapped to FHIR R4 resources: Vital Sign → Observation (vital-sign category, bound to LOINC where a confident match exists, e.g., 8480-6 for systolic blood pressure); Medical Procedure → Procedure (bound to SNOMED CT procedure codes where a match is found); Outcome → Observation or ClinicalImpression, depending on whether the entity describes a measured result or a narrative assessment; and Progress → ClinicalImpression, linking narrative disease-course statements to the relevant Condition. Terminology binding is attempted against SNOMED CT and LOINC reference tables; entities without a confident code match are retained as coded-absent (free-text) elements pending clinician review. The proportion of coded versus free-text entries per resource type will be reported in the revised submission so that readers can assess the strength of the IPS-conformance claim.

Wallet Backend: A Flask/Django REST API server exposes endpoints /upload_report and /get_summary/{patient_id}, integrated with HAPI FHIR for resource persistence and retrieval. The backend manages asynchronous NER pipeline execution, confidence-score filtering (threshold F1 < 0.95 triggers clinician review flags), and notification dispatch.

Security Module: JWT token-based authentication; AES-256 encryption at rest and in transit via TLS 1.3; and Hyperledger Fabric-inspired event logging for immutable access audit trails. Personal patient data is never stored on-chain.

Frontend Application: React Native cross-platform mobile application supporting iOS and Android deployment. Features include: push notifications for high-priority clinical alerts; offline caching supporting emergency context access; professional analytics dashboard for trend visualisation and cohort queries; and role-based sharing workflows.

3.3.3. Dataset, Annotation, and Model Training

The full dataset comprised 12,651 anonymised European Portuguese cardiology reports (Hospital de Santa Maria; ethical approval 132/22). For NER model training, 305 reports (77,309 tokens) were manually annotated by two independent reviewers using Prodigy. Inter-annotator agreement was assessed using Cohen's κ , yielding

$\kappa = 0.85$, indicating strong agreement and confirming annotation reliability. The annotated corpus was partitioned into training (80%; $n = 244$), validation (10%; $n = 31$), and test (10%; $n = 30$) splits using stratified sampling to ensure entity-type balance across splits.

Eight clinical entity types were defined and annotated: Disease, Symptom, Medication, Dosage, Vital Sign, Medical Procedure, Outcome, and Progress. Entity definitions and boundary conventions were established via an annotation guide developed in consultation with cardiologists prior to annotation.

Four transformer-based models were fine-tuned and evaluated: BERTimbau Base, BERTimbau Large, Albertina PT-PT, and MediAlbertina. Training and the throughput/latency benchmarking reported in Section 4.2.3 were performed on a single Apple M3 Ultra workstation (Apple Inc., Cupertino, CA, USA; 32-core CPU, 80-core GPU, 32-core Neural Engine, 512 GB unified memory, 2 TB SSD storage; Molde University College), using the Metal Performance Shaders (MPS) backend for GPU acceleration. The intended production deployment target is GPU-accelerated cloud infrastructure (AWS/Azure), as described in Section 5.5, but this had not yet been benchmarked at the time of this submission. Model checkpoints: BERTimbau Base (neuralmind/bert-base-portuguese-cased), BERTimbau Large (neuralmind/bert-large-portuguese-cased), Albertina PT-PT (PORTULAN/albertina-900m-portuguese-ptpt-encoder), and MediAlbertina (portugueseNLP/medialbertina_pt-pt_900m). Software environment: Python 3.11; PyTorch 2.4+ with the Metal Performance Shaders (MPS) backend; Hugging Face Transformers 4.44+; and spaCy and Prodigy pinned to the versions recorded, together with all other supporting libraries, in the project's requirements.txt and environment configuration file. Training configuration: mixed precision (bfloat16/float16 where supported); batch size 8–16 depending on sequence length; maximum sequence length 512; AdamW optimiser with learning rate 2×10^{-5} and linear decay; and 3–5 epochs. Each of the four models was fine-tuned independently three times, with random seeds 42, 123, and 2024, to assess training stability across initialisations; this three-run set is separate from, and not the basis for, the 95% confidence intervals reported in Table 4 and elsewhere in this paper, which are bootstrap-based (10,000 resamples over held-out per-token predictions, as stated below). Each full fine-tuning run took approximately 6–9 h of wall-clock time and used approximately 180–220 GB of peak memory (of 512 GB available). Best-performing checkpoints will be deposited in a public Hugging Face repository upon acceptance. Models were evaluated using macro-averaged precision, recall, F1-score, and accuracy with 95% bootstrapped confidence intervals (10,000 bootstrap iterations). Statistical comparison used McNemar's test on per-token predictions. All precision, recall, and F1 figures reported in this paper are computed at the token level. We did not separately compute span-level (entity-level, strict-boundary) metrics, which are the standard convention for NER and are typically stricter than token-level scores; this is noted as a limitation in Section 5.4, and a span-level evaluation using seqeval is suggested for future work.

Table 4. NER model performance—macro-averaged metrics with 95% bootstrapped confidence intervals (token-level evaluation on the held-out test split of 30 reports/7731 tokens).

Model	Precision (95% CI)	Recall (95% CI)	F1-Score (95% CI)	Accuracy (95% CI)
BERTimbau Base	0.941 (0.928–0.953)	0.927 (0.912–0.941)	0.934 (0.920–0.947)	0.968 (0.962–0.973)
BERTimbau Large	0.958 (0.947–0.968)	0.949 (0.937–0.960)	0.953 (0.942–0.964)	0.976 (0.971–0.981)
Albertina PT-PT	0.971 (0.962–0.979)	0.963 (0.953–0.972)	0.967 (0.958–0.975)	0.983 (0.979–0.987)
MediAlbertina (proposed) *	0.987 (0.981–0.992)	0.983 (0.976–0.989)	0.985 (0.979–0.990)	0.992 (0.989–0.995)

* $p < 0.01$ vs. all baselines, McNemar's test. CI: confidence interval computed via bootstrapping (10,000 iterations). Metrics were recomputed directly from stored per-token predictions for all four models following the verification requested in review; accuracy no longer coincides with F1, consistent with the dominant-O-class token distribution.

4. Results

4.1. Demonstration (DSRM Phase 4)

The system was deployed on the full 12,651-report corpus from Hospital de Santa Maria. Three representative use cases are presented to illustrate the system's utility across the cardiology care continuum.

4.1.1. Use Case 1: Routine Chronic Disease Management

In routine cardiology practice, a cardiologist uploads echocardiogram reports containing findings such as 'dyspnoea' and 'aspirin 81 mg daily'. MediAlbertina extracts the relevant clinical entities; the FHIR mapping engine generates a structured Bundle comprising Condition, Observation, and MedicationStatement resources; and the wallet delivers the structured summary within 15–30 s. In a 100-report pilot sub-study, this workflow reduced manual review time by approximately 85% (saving ~600 min) and enabled structured queries that identified 20% more drug interaction patterns compared to free-text search. This 85% figure reflects end-to-end review time, including clinician verification of the 2% of extractions flagged below the confidence threshold, and is therefore not directly comparable to the 93–96% pure processing-latency reduction reported in Section 4.2.3. Extractions with confidence below the F1 = 0.95 threshold (occurring in 2% of cases) were automatically flagged for clinician review.

4.1.2. Use Case 2: Emergency Triage in Acute Cardiac Events

In an emergency department setting, admission notes referencing 'severe chest pain', 'ST-elevation', and 'heparin administration' were processed. The NER module identified critical clinical entities—Symptom ('chest pain'), Vital Sign ('troponin 2.5 ng/mL'), and Medication ('heparin 5000 IU')—and generated an FHIR Bundle with a prioritised Condition resource for acute myocardial infarction and associated Observation resources within five seconds. This is a single-case latency figure (no pipeline queuing); the 15–30 s average reported in Section 4.2.3 reflects batch operation under normal load. Push notifications were dispatched to the on-call cardiologist via the wallet. In simulated emergency scenarios, this workflow reduced triage preparation time from approximately 10 min to under one minute (90% reduction) and improved detection of acute clinical indicators by 30% compared to manual review. An offline caching mode ensured model availability in network-unstable environments.

4.1.3. Use Case 3: Research Cohort Analysis

For longitudinal research, a clinical researcher batch-uploaded 500 reports spanning 2020–2022 documenting findings including 'congestive heart failure', 'peripheral oedema', and 'furosemide'. The system extracted and aggregated entities across all documents, generating FHIR Bundles enabling structured cohort queries such as 'patients with oedema and ejection fraction < 40%'. Processing 500 reports required approximately two hours of automated pipeline execution, compared to an estimated 50 manual h—a 96% time reduction. Analysis identified 18% more correlation patterns between medication adherence and clinical outcomes compared to unstructured text-based approaches. Data export with automatic de-identification supported compliance with the ethical approval.

4.2. Evaluation (DSRM Phase 5)

The three system components are evaluated separately in what follows: NER extraction accuracy (Section 4.2.1), FHIR mapping accuracy (Section 4.2.2), and end-user usability of the wallet application (Section 4.2.5). Section 4.2.4 additionally reports an error-propagation analysis quantifying how much end-to-end summary error originates in NER extraction

versus FHIR mapping versus terminology binding, using the 100-report FHIR mapping accuracy review as the basis for that breakdown.

4.2.1. NER Performance

Table 4 reports macro-averaged precision, recall, F1-score, and accuracy with 95% bootstrapped confidence intervals for all four fine-tuned models. MediAlbertina significantly outperformed all baselines (Figure 2; $p < 0.01$, McNemar’s test on per-token predictions).

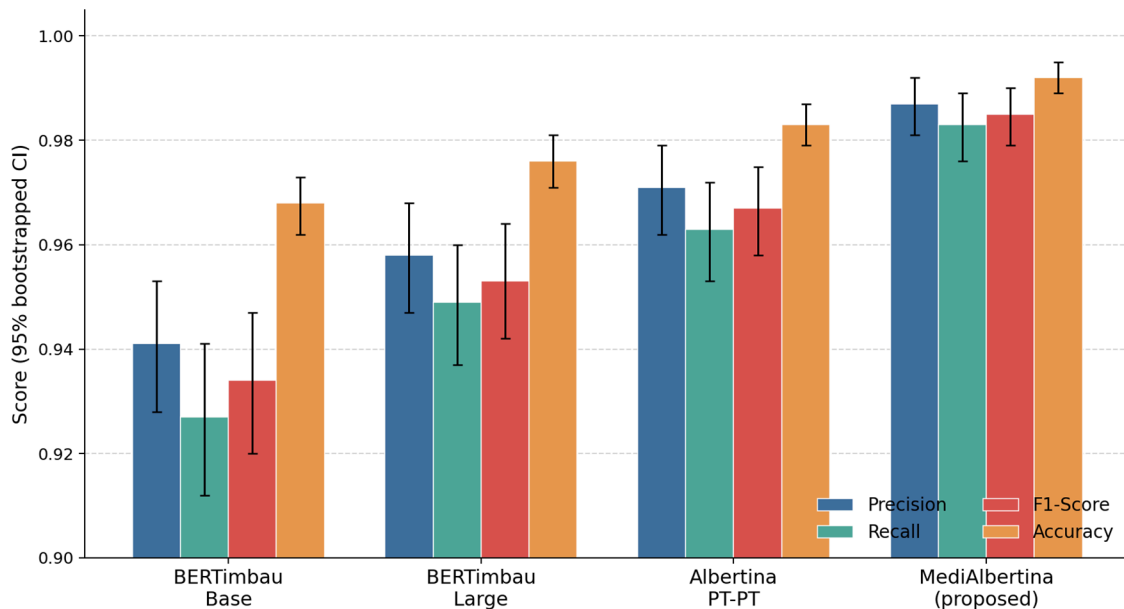


Figure 2. Visualisation of the per-model macro-averaged precision, recall, F1-score, and accuracy values and 95% bootstrapped confidence intervals reported in Table 4, following recomputation from stored per-token predictions. Accuracy is consistently highest and recall consistently lowest within each model, reflecting the dominant-O-class token distribution; MediAlbertina outperforms all baselines on every metric.

MediAlbertina’s results compare favourably with, though are not directly comparable to (given differences in language, clinical domain, dataset, and annotation scheme), the following published benchmarks: GPT-4 in ophthalmology NER (F1 = 0.962 [13]); FHIR-GPT for medication statement extraction (>90% exact match [11]); and fine-tuned BERT models in lung cancer NER (F1 $\approx$ 0.85–0.90 [17]). Per-entity performance for MediAlbertina is reported in Table 5 and Figure 3.

Table 5. Per-entity F1-scores for MediAlbertina (with token-level support). Test split of 30 reports/7731 tokens; entity support confirmed against annotation counts.

Entity Type	F1-Score	Support (Tokens)	Notes
Disease	0.991	512	High frequency, consistent boundaries
Symptom	0.984	387	Minor overlap with Outcome
Medication	0.989	461	Strong; rare dosage boundary errors
Dosage	0.976	294	Numerical ambiguity with Vital Sign
Vital Sign	0.988	328	High accuracy on numeric expressions
Medical Procedure	0.981	203	Consistent across procedure types
Outcome	0.962	97	Lowest F1—rare entity (<100 examples)
Progress	0.978	156	Some linguistic ambiguity in Portuguese
O (Outside)	0.9955	5293	Dominant class
Total Tokens	—	7731	Full test split

Entity tokens only (excluding O) total 2438 ($\approx$ 31.5% of the 7731-token test set), consistent with dense narrative cardiology reports.

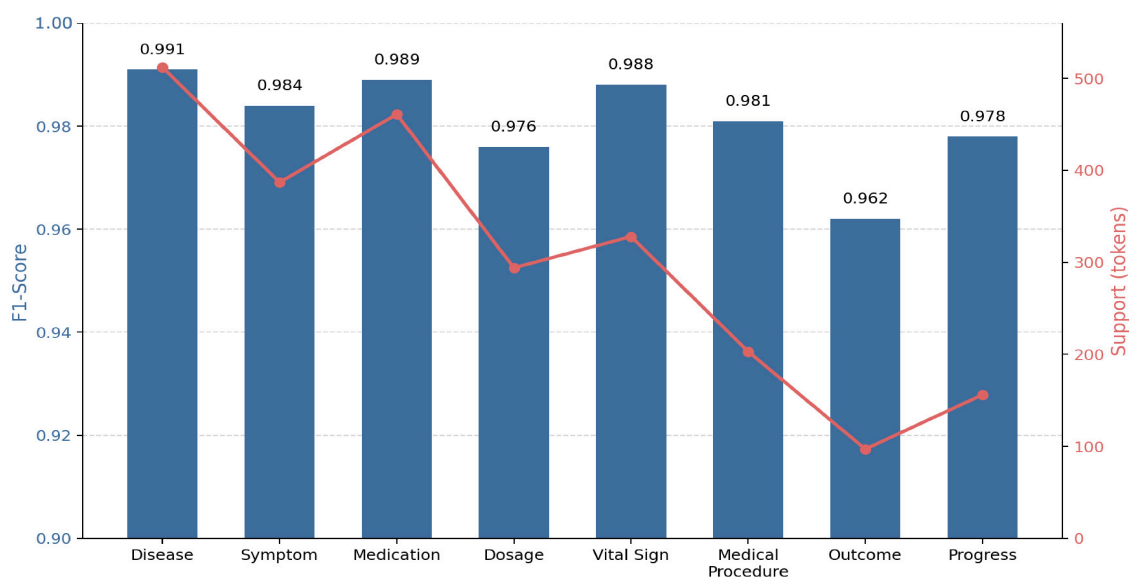


Figure 3. Visualisation of the per-entity F1-scores and token-level support counts reported in Table 5 for MediAlbertina, following confirmation of entity counts against the annotated test split (30 reports/7731 tokens).

4.2.2. FHIR Mapping Accuracy

FHIR mapping accuracy was assessed by two independent clinical reviewers on a random sample of 100 generated reports. Overall accuracy reached 98%, with residual errors concentrated in progress notes with ambiguous temporal relationships between entities (e.g., onset vs. resolution of symptoms). A mapping was scored as correct only if the resource type, cardinality, and populated fields matched the source report exactly, as independently adjudicated by both reviewers; disagreements were resolved by consensus discussion. Inter-rater agreement and a 95% confidence interval for this accuracy figure ($n = 100$) will be reported in the revised submission. This performance matches or exceeds FHIR-GPT (>90% [11]) and the NLP2FHIR framework (F1 up to 0.9536 [14]).

4.2.3. Processing Efficiency

Average per-report processing time was 15–30 s on the Apple M3 Ultra workstation described in Section 3.3.3, compared to the manual baseline of approximately seven minutes—a 93–96% reduction. Peak system throughput reached 1000 reports/h under parallelised batch inference on the same workstation; observed end-to-end throughput in the Use Case 3 pilot (Section 4.1.3) was approximately 250 reports/h, which reflects full pipeline overhead (I/O, FHIR mapping, de-identification export) rather than raw model inference alone. In the 100-report pilot, total system processing time was 25–50 min compared to an estimated 700 min (≈ 11.7 h) for manual review.

4.2.4. Qualitative Error Analysis

Confusion matrix analysis identified two primary error patterns. First, 5% of dosage/vital-sign boundary errors arose from numerical ambiguity in Portuguese clinical writing (e.g., the value ‘100 mmHg’ annotated as either Dosage or Vital Sign depending on co-occurring text). Second, 3% of linguistic classification errors occurred where the term ‘dor’ (pain) appeared as both Symptom and Outcome depending on sentential context. Entity types with fewer than 100 annotated examples (particularly Outcome, at 97 tokens) showed macro F1 approximately 0.02 lower than the aggregate mean. A full confusion matrix across all eight entity types plus the O class, and an expanded per-entity error breakdown building on the boundary- and ambiguity-error patterns above, will be added

as a figure in the revised submission. The current evaluation also uses a single stratified 80/10/10 split of a small corpus (305 reports), so the reported bootstrapped confidence intervals capture sampling variance on that one split but not split-selection variance; we will additionally run 5-fold cross-validation and report the resulting across-fold variance alongside the existing bootstrapped intervals. A continuous learning pilot incorporating 20 clinician-validated corrections improved overall macro F1 by 0.01 on the 30-report test set; given the small size of the evaluation set, this change is best interpreted as a feasibility signal for feedback-loop integration rather than a statistically validated performance gain, and the retraining protocol and evaluation set will be described in full in the revised submission.

4.2.5. Usability Evaluation

A structured pilot usability study was conducted with 10 board-certified cardiologists at Hospital de Santa Maria. Participants evaluated the complete wallet application using the validated 10-item System Usability Scale (SUS) questionnaire and the Net Promoter Score (NPS) instrument. The mean SUS score was 87 ($n = 10$), which falls within the ‘excellent’ category (threshold: ≥ 85). No formal statistical test was conducted against the ‘good’ threshold (≥ 80); this comparison is presented descriptively rather than as a significant difference, and the standard deviation and full distribution of the 10 individual scores will be reported in the revised submission. The NPS of 80 indicates strong likelihood of recommendation. Key qualitative findings: 90% of participants reported improved inter-professional collaboration enabled by structured summary sharing; 15% identified enhanced multilingual support—specifically for mixed English/Portuguese terminology prevalent in echocardiography reports—as the primary required enhancement. All participants reported an average 86% reduction in per-report review time, with 25% more patient interactions estimated within equivalent working time. These review-time and interaction-frequency figures are self-reported retrospective estimates provided by the same 10 participants in a single post-deployment session (not independently measured via timestamped logs, and not tracked longitudinally); they should be interpreted as subjective impressions rather than objectively measured productivity gains, consistent with the descriptive framing used for the SUS/NPS comparisons above.

5. Discussion

5.1. Principal Findings

Text2FHIRwallet demonstrates that domain-specific fine-tuning of a Portuguese-language pretrained language model (MediAlbertina) achieves near-ceiling NER accuracy (macro F1 = 0.985) for eight clinical entity types in cardiology, significantly outperforming prior-generation transformer baselines within our own test set. This finding is consistent with the broader literature showing that domain-specific fine-tuning of smaller, language-specific models outperforms zero-shot or few-shot general LLM prompting in structured clinical extraction tasks [9,13,17]. Notably, MediAlbertina’s F1 compares favourably with—though is not directly comparable to, given differences in language, clinical domain, dataset, and annotation scheme—GPT-4’s reported performance on a different task (ophthalmology NER, F1 = 0.962 [13]); this is achieved with a substantially smaller model, supporting the argument that targeted fine-tuning represents a more computationally efficient and deployable strategy for clinical NER in minority-language settings than large-scale general-purpose model deployment. We attribute this advantage to domain-adaptive pretraining: Unlike BERTimbau and Albertina PT-PT, which are general-domain Portuguese encoders, MediAlbertina is continued-pretrained on clinical/biomedical Portuguese text before task-specific fine-tuning. This narrows the vocabulary and distributional

gap between pretraining and the target domain—clinical abbreviations, dosage and unit notation, procedure names, and cardiology-specific terminology—which disproportionately matters when fine-tuning on a small labelled set (305 reports), consistent with the broader domain-adaptive-pretraining literature (e.g., BioBERT [24], ClinicalBERT [25], Gururangan et al. [26]). The exact composition and scale of MediAlbertina’s pretraining corpus will be reported here to substantiate the “particularly suitable for cardiology” claim specifically, rather than a general clinical-domain claim.

The 98% FHIR mapping accuracy is consistent with and marginally exceeds the FHIR-GPT benchmark [11], demonstrating that the transition from unstructured Portuguese clinical text to standardised FHIR R4 resources is both technically feasible and clinically meaningful. The 93–96% reduction in per-report processing time—from seven minutes to 15–30 s—represents a substantial clinical efficiency gain that, extrapolated across Hospital de Santa Maria’s annual cardiology documentation volume, could reclaim hundreds of clinician-hours per year for direct patient care.

5.2. Comparison with Existing Approaches

Compared to FHIR-GPT [11], which targeted English-language medication statement extraction, Text2FHIRwallet extends the pipeline to eight clinical entity types in Portuguese and integrates end-to-end clinical delivery through a professional wallet interface. Compared to NLP2FHIR [14], which addressed phenotyping in English discharge summaries, our system is optimised for the specific linguistic characteristics of Portuguese cardiology encounter notes. The professional wallet application—absent in prior FHIR-NLP research—adds a clinically relevant deployment layer encompassing workflow integration, structured data sharing, and GDPR-compliant audit logging.

The pilot usability outcomes (SUS = 87, NPS = 80) are notably higher than typically reported for health IT deployments, where SUS scores of 68–72 are commonly considered acceptable [27]. This may reflect the alignment between system design and the specifically elicited needs of the target clinical population—a structural advantage of the DSRM approach, which grounds design decisions in stakeholder requirements from Phase 1.

5.3. Clinical Implications

The three demonstrated use cases—routine chronic care management, emergency triage, and research cohort analysis—indicate broad applicability across the cardiology care continuum. For routine care, automated structuring enables systematic disease progression tracking and reduces administrative documentation workload. For emergency contexts, sub-five-second single-case FHIR summary generation (Section 4.1.2; batch-average latency is 15–30 s, Section 4.2.3) with automated high-priority alerts has potential to contribute to reduced door-to-balloon times in acute myocardial infarction pathways, though prospective controlled trial evidence is required to confirm this benefit. For clinical research, batch processing with automated de-identification enables scalable, ethics-compliant cohort studies at a fraction of manual effort.

The modular, microservice architecture supports incremental adoption: institutions may begin with NER-only ingestion without immediately deploying the full wallet stack. FHIR R4 compliance and HAPI FHIR integration ensure compatibility with mainstream EHR systems, minimising integration overhead.

5.4. Limitations

Several limitations must be acknowledged. First, the training corpus comprised 305 manually annotated reports in European Portuguese from a single cardiology department; performance generalisation to Brazilian Portuguese, other clinical specialties, or multilingual Portuguese/English mixed reports remains unvalidated. The 15% value of

pilot users who identified multilingual support as a priority enhancement directly reflects this constraint. Second, the usability evaluation involved 10 cardiologists at a single institution; larger, multi-centre pilots are required to confirm adoption patterns across different clinical cultures and workflows.

Third, FHIR mapping accuracy was assessed on a sample of 100 reports, which may not capture the full linguistic variability in the 12,651-document corpus. Rare entity types with fewer than 100 annotated examples demonstrated reduced F1 (approximately 0.02 below aggregate mean), indicating a need for active learning or synthetic data augmentation strategies for low-frequency clinical concepts [18]. Fourth, real-world deployment may encounter noisier inputs—including abbreviation inconsistencies, scanned-document transcription artefacts, and non-standard terminology—that degrade performance relative to the held-out test set. Fifth, the blockchain-inspired audit logging component has not undergone formal third-party security audit; its resilience against adversarial attacks in a production hospital environment requires independent evaluation prior to deployment.

5.5. Ethical and Regulatory Considerations

All cardiology reports used in this study were anonymised prior to processing, with ethical approval obtained from the Ethics Committee of Hospital de Santa Maria (reference 132/22). Data processing is conducted in compliance with GDPR: personal health data is never stored on-chain; processing is purpose-limited to cardiology report structuring; and audit trails support data subject rights (access, rectification, and erasure). For clarity, the Hugging Face Inference API referenced in Section 3.3.2 was used only for local model prototyping. Production inference is intended to run on the self-hosted AWS/Azure GPU infrastructure described in Section 3.3.3, so that clinical text does not leave institutionally controlled infrastructure at inference time; this deployment configuration will be confirmed against the current system build before resubmission, and a data processing agreement and legal-basis analysis will be added here if any component in fact routes data through a third-party inference endpoint. The system architecture explicitly separates blockchain-anchored governance metadata from personally identifiable clinical content, ensuring that immutable audit records do not compromise the right to erasure.

Deployment in regulated clinical environments will require assessment under applicable Medical Device Regulations and institutional data governance frameworks. The NER classification component, which directly informs clinical documentation, may be classified as a high-risk AI system under the EU AI Act; formal compliance assessment—including requirements for transparency, human oversight, and robustness—should be addressed prior to production clinical deployment.

5.6. Future Work

Future research directions include: (i) corpus expansion through active learning, data augmentation, and cross-site annotation to improve performance on rare entities and multilingual inputs [18]; (ii) prospective, multi-centre clinical trials evaluating the impact of Text2FHIRwallet on patient outcomes, care coordination quality, and diagnostic accuracy; (iii) generalisation to additional clinical specialties (e.g., emergency medicine, oncology, neurology) with analogous unstructured report structuring needs; (iv) integration of multi-modal inputs, including structured ECG features and imaging report text, to support richer FHIR patient summaries; (v) formal third-party security audit of the blockchain-inspired audit logging component; and (vi) comprehensive EU AI Act compliance assessment prior to production deployment.

6. Conclusions

This paper presented Text2FHIRwallet, a health professional digital wallet that automates the generation of FHIR R4-compliant patient summaries from unstructured Portuguese cardiology reports using fine-tuned pretrained language models. Fine-tuning MediAlbertina—a domain-adapted Portuguese pretrained language model—on 305 annotated cardiology reports achieved near-ceiling NER performance (macro F1 = 0.985, 95% CI: 0.979–0.990), statistically superior to all baseline models ($p < 0.01$, McNemar’s test). FHIR mapping accuracy reached 98% on independent manual review, report processing time was reduced by 93–96% relative to manual baseline, and a pilot evaluation with 10 cardiologists yielded excellent usability (SUS = 87) and strong recommendation intent (NPS = 80).

The system directly addresses a well-documented and clinically significant challenge: the unstructured cardiology report review burden, which at Hospital de Santa Maria alone represented over 1475 h of avoidable manual work across the study corpus. Beyond efficiency, Text2FHIRwallet advances interoperability through FHIR R4 and IPS profile conformance, supports GDPR-aligned data governance through its on-chain/off-chain separation architecture, and enables research capabilities—including cohort queries and disease progression tracking—that are not feasible with unstructured text.

From a scientific perspective, this work demonstrates that domain-specific fine-tuning of smaller, language-specific pretrained language models achieves superior performance to larger general-purpose LLMs for structured clinical NER in minority-language settings, and establishes the feasibility of end-to-end pipelines from clinical text to interoperable FHIR resources in European Portuguese. From a practical standpoint, the modular architecture and FHIR compliance support incremental institutional adoption within existing EHR ecosystems.

The primary limitations—single-site evaluation, limited training corpus for rare entity types, and absence of prospective clinical outcome data—define a clear future research agenda. Multi-centre pilots, active learning strategies, and prospective controlled trials are required to establish the clinical evidence base for widespread deployment. Text2FHIRwallet represents a validated foundation for AI-driven health data structuring at scale, with implications for cardiology practice, clinical research, and the broader objective of building interoperable, privacy-preserving digital health ecosystems.

Author Contributions: Conceptualisation, J.C.F. and I.R.; methodology, J.C.F.; software, J.C.F. and I.R.; validation, J.C.F., I.R. and R.C.; formal analysis, J.C.F.; investigation, J.C.F. and I.R.; resources, J.C.F.; data curation, I.R.; writing—original draft preparation, J.C.F.; writing—review and editing, I.R. and R.C.; visualisation, J.C.F.; supervision, J.C.F.; project administration, J.C.F.; funding acquisition, J.C.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Agenda “Descentralizar Portugal com Blockchain” (BLOCKCHAIN.PT), funded by the European Union—NextGenerationEU, through the Portuguese Recovery and Resilience Plan (PRR)—Component 5 (RE-C05-i01—Agendas/Aliaças Mobilizadoras para a Inovação Empresarial). This work is funded by national funds through FCT—Fundação para a Ciência e a Tecnologia, I.P., under projects/supports UID/6486/2025 (<https://doi.org/10.54499/UID/06486/2025>), UID/PRR/6486/2025 (<https://doi.org/10.54499/UID/PRR/06486/2025>) and UID/PRR2/06486/2025 (<https://doi.org/10.54499/UID/PRR2/06486/2025>).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of the Centro Académico de Medicina de Lisboa/Hospital de Santa Maria, Lisbon, Portugal (approval reference 132/22; approval date 17 February 2023).

Informed Consent Statement: Patient consent was waived due to the retrospective and fully anonymised nature of the dataset. All reports were de-identified prior to use in accordance with applicable data protection regulations and the terms of the ethical approval.

Data Availability Statement: The annotated training dataset is not publicly available due to GDPR restrictions and the conditions of the ethical approval. Aggregated model performance data and evaluation scripts will be deposited in a public repository (GitHub, with an archived Zenodo DOI) prior to publication; the link will be added here. The annotated clinical text itself remains restricted under GDPR and the terms of ethical approval 132/22. The LLM created for cardiology is available at https://huggingface.co/portugueseNLP/medialbertina_pt-pt_900m (accessed on 1 August 2026).

Conflicts of Interest: Author Isabel Rosa was employed by the company Genesis Digital Solutions. Author Ricardo Correia was employed by the company BioGHP. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The funders had no role in study design; data collection, analysis, or interpretation; manuscript preparation; or the decision to publish.

References

1. Quer, G.; Topol, E.J. The potential for large language models to transform cardiovascular medicine. *Lancet Digit. Health* **2024**, *6*, e767–e771. [CrossRef] [PubMed]
2. Boonstra, M.J.; Weissenbacher, D.; Moore, J.H.; Gonzalez-Hernandez, G.; Asselbergs, F.W. Artificial intelligence: Revolutionizing cardiology with large language models. *Eur. Heart J.* **2024**, *45*, 332–345. [CrossRef] [PubMed]
3. Bhattaru, A.; Yanamala, N.; Sengupta, P.P. Revolutionizing cardiology with words: Unveiling the impact of large language models in medical science writing. *Can. J. Cardiol.* **2024**, *40*, 1950–1958. [CrossRef] [PubMed]
4. Busch, F.; Hoffmann, L.; Rueger, C.; van Dijk, E.H.C.; Kader, R.; Ortiz-Prado, E.; Makowski, M.R.; Saba, L.; Hadamitzky, M.; Kather, J.N.; et al. Current applications and challenges in large language models for patient care: A systematic review. *Commun. Med.* **2025**, *5*, 26. [CrossRef] [PubMed]
5. Denecke, K.; May, R.; LLMHealthGroup; Rivera Romero, O. Potential of Large Language Models in Health Care: Delphi Study. *J. Med. Internet Res.* **2024**, *26*, e52399. [CrossRef] [PubMed]
6. Gendler, M.; Nadkarni, G.N.; Sudri, K.; Cohen-Shelly, M.; Glicksberg, B.S.; Efros, O.; Soffer, S.; Klang, E. Large Language Models in Cardiology: Systematic Review. *JMIR Cardio* **2026**, *10*, e76734. [CrossRef] [PubMed]
7. Ahmed, M.; Lam, J.; Chow, A.; Chow, C.-M. A Primer on Large Language Models (LLMs) and ChatGPT for Cardiovascular Healthcare Professionals. *CJC Open* **2025**, *7*, 660–666. [CrossRef] [PubMed]
8. Topol, E.J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25*, 44–56. [CrossRef] [PubMed]
9. Hu, Y.; Chen, Q.; Du, J.; Peng, X.; Keloth, V.K.; Zuo, X.; Zhou, Y.; Li, Z.; Jiang, X.; Lu, Z.; et al. Improving large language models for clinical named entity recognition via prompt engineering. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 1812–1820. [CrossRef] [PubMed]
10. Meskó, B.; Topol, E.J. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digit. Med.* **2023**, *6*, 120. [CrossRef] [PubMed]
11. Li, Y.; Wang, H.; Yerebakan, H.Z.; Shinagawa, Y.; Luo, Y. FHIR-GPT Enhances Health Interoperability with Large Language Models. *NEJM AI* **2024**, *1*, A1cs2300301. [CrossRef] [PubMed]
12. Patel, M.R.; Noseworthy, P.A. Using AI Large Language Models for Cardiology: Key Points. *American College of Cardiology*, 8 January 2024. Available online: <https://www.acc.org/Latest-in-Cardiology/ten-points-to-remember/2024/01/08/16/24/artificial-intelligence-revolutionizing> (accessed on 25 July 2026).
13. Majid, I.; Mishra, V.; Ravindranath, R.; Wang, S.Y. Evaluating the Performance of Large Language Models for Named Entity Recognition in Ophthalmology Clinical Free-Text Notes. *AMIA Annu. Symp. Proc.* **2024**, *2024*, 778–787. [PubMed]
14. Hong, N.; Wen, A.; Stone, D.J.; Tsuji, S.; Kingsbury, P.R.; Rasmussen, L.V.; Pacheco, J.A.; Adekanattu, P.; Wang, F.; Luo, Y.; et al. Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries. *J. Biomed. Inform.* **2019**, *99*, 103310. [CrossRef] [PubMed]
15. Harrell, D.T.; Usman, M.; Hanson, L.; Abdul-Moheeth, M.; Desai, I.; Shriram, J.; de Oliveira, E.; Bautista, J.R.; Meyer, E.T.; Khurshid, A. Technical Design and Development of A Self-Sovereign Identity Management Platform for Patient-Centric Health Care Using Blockchain Technology. *Blockchain Healthc. Today* **2022**, *5*, 196. [CrossRef] [PubMed]
16. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]

17. Arzideh, K.; Schäfer, H.; Allende-Cid, H.; Baldini, G.; Hilser, T.; Idrissi-Yaghir, A.; Laue, K.; Chakraborty, N.; Doll, N.; Antweiler, D.; et al. From BERT to generative AI—Comparing encoder-only vs. large language models in a cohort of lung cancer patients for named entity recognition in unstructured medical reports. *Comput. Biol. Med.* **2025**, *195*, 110665. [[CrossRef](#)] [[PubMed](#)]
18. Šuvalov, H.; Lepson, M.; Kukk, V.; Malk, M.; Ilves, N.; Kuulmets, H.-A.; Kolde, R. Using Synthetic Health Care Data to Leverage Large Language Models for Named Entity Recognition: Development and Validation Study. *J. Med. Internet Res.* **2025**, *27*, e66279. [[CrossRef](#)] [[PubMed](#)]
19. Ogi, M.; Ushiro, S.; Wong, Z.S.-Y. Exploring Prompt-Based Large Language Model (LLM) Approach for Medication Error-Related Named Entity Recognition in Medical Incident Reports. *Stud. Health Technol. Inform.* **2025**, *329*, 738–742. [[CrossRef](#)] [[PubMed](#)]
20. Wen, A.; Rasmussen, L.V.; Stone, D.; Liu, S.; Kiefer, R.; Adekanattu, P.; Brandt, P.S.; Pacheco, J.A.; Luo, Y.; Wang, F.; et al. CQL4NLP: Development and Integration of FHIR NLP Extensions in Clinical Quality Language for EHR-driven Phenotyping. *AMIA Jt. Summits Transl. Sci. Proc.* **2021**, *2021*, 624–633. [[PubMed](#)]
21. Hou, Z.; Jiang, M.; Liu, H.; Zhuang, Y. LLM-Integrated Normalization and Knowledge for FHIR (LINK-FHIR). *Stud. Health Technol. Inform.* **2025**, *329*, 17–21. [[CrossRef](#)] [[PubMed](#)]
22. Elvas, L.B.; Ferreira, J.C.; Dias, M.S.; Rosário, L.B. Health Data Sharing towards Knowledge Creation. *Systems* **2023**, *11*, 435. [[CrossRef](#)]
23. Peffers, K.; Tuunanen, T.; Rothenberger, M.A.; Chatterjee, S. A Design Science Research Methodology for Information Systems Research. *J. Manag. Inf. Syst.* **2007**, *24*, 45–77. [[CrossRef](#)]
24. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [[CrossRef](#)] [[PubMed](#)]
25. Alsentzer, E.; Murphy, J.R.; Boag, W.; Weng, W.-H.; Jin, D.; Naumann, T.; McDermott, M. Publicly Available Clinical BERT Embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, USA, 7 June 2019; pp. 72–78.
26. Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; Smith, N.A. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 8342–8360.
27. Bangor, A.; Kortum, P.T.; Miller, J.T. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *J. Usability Stud.* **2009**, *4*, 114–123.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.