Kaufman, L. and P. J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, 1 edition.

Kingma, D. P. and J. Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representation (ICLR 2015)*.

Larson, S., A. Mahendran, J. J. Peper, C. Clarke, A. Lee, P. Hill, J. K. Kummerfeld, K. Leach, M. A. Laurenzano, L. Tang, and J. Mars. 2019. An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316. ACL.

Levina, E. and P. J. Bickel. 2004. Maximum Likelihood Estimation of Intrinsic Dimension. In *Advances in Neural Information Processing Systems (NIPS 2004)*, volume 17, pages 777–784. MIT Press.

Liu, X., A. Eshghi, P. Swietojanski, and V. Rieser. 2021. Benchmarking Natural Language Understanding Services for building Conversational Agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, pages 165–183. Springer.

Lloyd, S. P. 1982. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

Maheswaranathan, N., A. Williams, M. Golub, S. Ganguli, and D. Sussillo. 2019. Reverse engineering Recurrent Networks for Sentiment Classification Reveals Line Attractor Dynamics. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, volume 32, pages 15696–15705. Curran Associates.

Martelli, M. 1999. *Introduction to Discrete Dynamical Systems and Chaos*. Wiley-Interscience, 1st edition.

Ming, Y., S. Cao, R. Zhang, Z. Li, Y. Chen, Y. Song, and H. Qu. 2017. Understanding Hidden Memories of Recurrent Neural Networks. In *Proceedings of the 2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 13–24. IEEE.

Morcos, A. S., D. G. T. Barrett, N. C. Rabinowitz, and M. Botvinick. 2018. On the Importance of Single Directions for Generalization. In *Proceedings of the 6th International Conference on Learning Representation (ICLR 2018)*.

Niimi, Y., T. Oku, T. Nishimoto, and M. Araki. 2001. A Rule Based Approach to Extraction of Topics and Dialog Acts in a Spoken Dialog System. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, pages 2185–2188. ISCA.

Rastogi, A., X. Zang, S. Sunkara, R. Gupta, and P. Khaitan. 2020. Towards Scalable Multi-domain Conversational Agents: The Schema-Guided Dialogue Dataset. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI-20)*, volume 34, pages 8689–8696. AAAI Press.

Ravuri, S. and A. Stolcke. 2015. Recurrent Neural Network and LSTM Models for Lexical Utterance Classification. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech 2015)*, pages 135–139. ISCA.

Rousseeuw, P. J. 1987. Silhouettes: A graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Sanchez-Karhunen, E., J. F. Quesada-Moreno, and M. A. Gutiérrez-Naranjo. 2024. Interpretation of the Intent Detection Problem as Dynamics in a Low-Dimensional Space. In *ECAI 2024: 27th European Conference on Artificial Intelligence*, volume 392, pages 3693–3700. IOS Press.

Sanchez-Karhunen, E., J. F. Quesada-Moreno, and M. A. Gutiérrez-Naranjo. 2025. Bias in Intent Detection: A Dynamical Systems Perspective. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25)*, volume 294, pages 396–402. PMLR.

Sanchez-Karhunen, E., J. F. Quesada-Moreno, and M. A. Gutiérrez-Naranjo. 2026. Interpretability of the Intent Detection Problem: A New Approach. *The European Journal on Artificial Intelligence*, 0(0). To appear.

Steinley, D. 2004. Properties of the Hubert-Arable Adjusted Rand Index. *Psychological Methods*, 9(3):386–396.

Strobelt, H., S. Gehrmann, H. Pfister, and A. M. Rush. 2018. LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):667–676.

Sussillo, D. and O. Barak. 2013. Opening the Black Box: Low-Dimensional Dynamics in High-Dimensional Recurrent Neural Networks. *Neural Computation*, 25(3):626–649.

Tenenbaum, J. B., V. de Silva, and J. C. Langford. 2000. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323.

Tomašev, N. and M. Radovanović. 2016. Clustering Evaluation in High-Dimensional Data. In *Unsupervised Learning Algorithms*. Springer, pages 71–107.

Tur, G. and R. De Mori. 2011. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. John Wiley & Sons.

Procesamiento del Lenguaje Natural, Revista nº 76, marzo de 2026, pp. 39-52 recibido 26-11-2025 revisado 17-01-2026 aceptado 10-02-2026

Automating Early Disease Prediction Via Structured and Unstructured Clinical Data

Automatización de la predicción temprana de enfermedades mediante datos clínicos estructurados y no estructurados

Ane G. Domingo-Aldama[1], Marcos Merino Prado[1], Alain García Olea[2], Josu Goikoetxea[1], Koldo Gojenola[1], Aitziber Atutxa[1]
[1]University of the Basque Country (EHU), [2]Basurto University Hospital
ane.garciad@ehu.eus, mmerinoprado@gmail.com, alain.garciaolea@osakidetza.eus, josu.goikoetxea@ehu.eus, koldo.gojenola@ehu.eus, aitziber.atucha@ehu.eus

Abstract: This study presents a fully automated methodology for early prediction studies in clinical settings, leveraging information extracted from unstructured discharge reports. The proposed pipeline uses discharge reports to support the three main steps of early prediction: cohort selection, dataset generation, and outcome labeling. By processing discharge reports with natural language processing techniques, we can efficiently identify relevant patient cohorts, enrich structured datasets with additional clinical variables, and generate high-quality labels without manual intervention. This approach addresses the frequent issue of missing or incomplete data in codified electronic health records (EHR), capturing clinically relevant information that is often underrepresented. We evaluate the methodology in the context of predicting atrial fibrillation (AF) progression, showing that predictive models trained on datasets enriched with discharge report information achieve higher accuracy and correlation with true outcomes compared to models trained solely on structured EHR data, while also surpassing traditional clinical scores. These results demonstrate that automating the integration of unstructured clinical text can streamline early prediction studies, improve data quality, and enhance the reliability of predictive models for clinical decision-making.
Keywords: Early Prediction, Atrial Fibrillation Progression, Electronic Health Records, Natural Language Processing.

Resumen: Este estudio presenta una metodología totalmente automatizada para estudios de predicción temprana en entornos clínicos, aprovechando la información extraída de informes de alta hospitalaria no estructurados. El proceso propuesto utiliza los informes de alta para respaldar los tres pasos principales de la predicción temprana: selección de cohortes, generación de conjuntos de datos y etiquetado de resultados. Mediante el procesamiento de los informes de alta con técnicas de procesamiento del lenguaje natural, podemos identificar de manera eficiente las cohortes de pacientes relevantes, enriquecer los conjuntos de datos estructurados con variables clínicas adicionales y generar etiquetas de alta calidad sin intervención manual. Este enfoque aborda el principal problema de los registros médicos electrónicos (RME) codificados que son los datos faltantes o incompletos, capturando información clínicamente relevante que a menudo está infrarrepresentada. Evaluamos la metodología en el contexto de la predicción de la progresión de la fibrilación auricular (FA), demostrando que los modelos predictivos entrenados con conjuntos de datos enriquecidos con información de informes de alta logran una mayor precisión y correlación con los resultados reales en comparación con los modelos entrenados únicamente con datos estructurados de RME, al tiempo que superan las puntuaciones clínicas tradicionales. Estos resultados demuestran que la automatización de la integración de texto clínico no estructurado puede agilizar los estudios de predicción temprana, mejorar la calidad de los datos y aumentar la fiabilidad de los modelos predictivos para la toma de decisiones clínicas.
Palabras clave: Predicción Temprana, Progresión de la Fibrilación Auricular, Historias Clínicas Electrónicas, Procesamiento del Lenguaje Natural.

ISSN 1135-5948 DOI 10.26342/2026-76-3 ©2026 Sociedad Española para el Procesamiento del Lenguaje Natural

1 Introduction

Early diagnosis (ED) has garnered significant attention in both Artificial Intelligence (AI) and medicine due to its transformative potential in personalized medicine. By identifying a condition at its initial stages, clinicians can intervene sooner, prevent complications, and tailor treatments to each patient which leads to improved outcomes and reduced healthcare costs. The success of these ED models heavily depends on the diversity, volume, and granularity of the data available (Ng et al., 2016; Alzubi, Watzlaf, and Sheridan, 2021).

Modern AI systems for personalized and predictive medicine predominantly employ machine learning (ML) models that rely exclusively on structured, tabular data (including: demographic information, laboratory measurements, coded diagnoses, etc.). This reliance has led to an increasing dependence on structured electronic health records (EHRs) (Ristevski and Chen, 2018). EHRs, now central to healthcare systems worldwide, provide a rich repository of both structured and unstructured patient information, playing a crucial role in clinical decision-making (Holmes et al., 2021; Alzubi, Watzlaf, and Sheridan, 2021).

Despite the rich information contained in structured EHRs, they present notable challenges, primarily stemming from the manual annotation and conversion of medical data into standardized structured and discrete fields. This process is prone to errors, missing values, and inconsistencies, which can significantly impact data reliability and the overall quality of the predictive models built upon this information (Garcia Olea et al., 2021; Botsis et al., 2010; Alzubi, Watzlaf, and Sheridan, 2021).

Efforts to standardize medical coding with systems like ICD-10, OPCS, and SNOMED have mitigated some challenges. However, no universal guidelines exist regarding the level of detail required for clinical documentation. Consequently, not all information relevant to specific prediction tasks is captured in structured EHRs, often necessitating extra manual annotation (a time-consuming process that introduces variability and potential human error).

Furthermore, ED studies require a cohort selection process to identify patients suitable for the task. This step typically depends on structured EHR data or, when key information is missing, on manual review of clinical records. As a result, the process becomes labor intensive, slows down research workflows, and introduces variability and potential human error (Chen et al., 2025; Jin et al., 2024).

In this context, the present study introduces a methodology that automates the key steps required for early prediction tasks, including cohort selection, dataset generation, and patient labeling. The approach combines structured tabular data derived from EHRs with semi-structured discharge reports in free-text format, processed through a natural language processing (NLP) pipeline. This methodology reduces the need for manual annotation while simultaneously improving the quality and completeness of structured tabular data by leveraging information extracted from discharge reports to mitigate missing values or codification errors. The primary contribution of this work lies in the design of this end-to-end methodology and data generation pipeline, rather than in proposing a novel predictive modeling architecture.

Specifically, we focus on the progression of atrial fibrillation (AF) within one month to two years after the initial episode (see Figure 1). The early prediction of AF progression in this time-window helps arrhythmia specialists determine whether rhythm control therapies are appropriate for a patient based on their individual risk. In this study, AF progression encompasses all atrial fibrillation subtypes: paroxysmal, persistent, and permanent[1]. All methodological steps are evaluated in this context, and the resulting predictive models are compared against traditional clinical scores for AF progression.

To this end, the research is guided by the following core questions:

- **RQ1**: Can the combination of structured EHR data and information extracted from discharge reports enhance and automate the development of ED models?
- **RQ2**: Can free-text discharge reports processed with NLP techniques improve the quality and completeness of structured tabular data derived from EHRs?

[1]For paroxysmal AF, progression is defined as the recurrence of the arrhythmia, whereas for persistent and permanent AF, it corresponds to the ongoing presence of the condition.

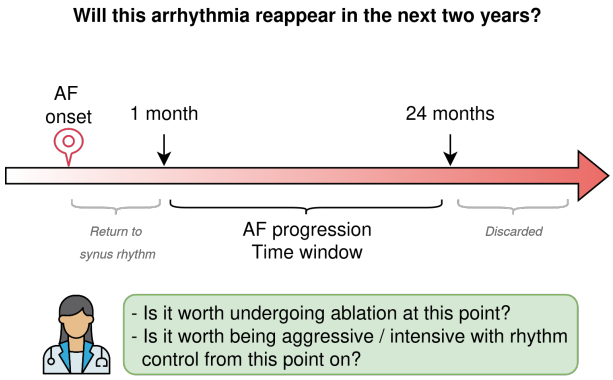

Figure 1: **Clinical scenario of AF progression.**

- **RQ3**: Can automatically generated silver annotations achieve performance comparable to gold-standard annotations while significantly reducing manual effort?
- **RQ4**: Does the proposed methodology produce predictive models that outperform existing clinical scores for AF progression prediction?

2 Related Work

Disease prediction (Sharma et al., 2022; Xie, Yu, and Lv, 2021; Yu et al., 2023; Mishra and Tarar, 2020) has attracted considerable attention in both the fields of AI and medicine due to its potential to revolutionize ED and personalized medicine (Awwalu et al., 2015; Shah, 2018; Schork, 2019; Ullah, Akbar, and Yannarelli, 2020). To carry out these studies three steps are necessary: cohort selection, dataset generation and labeling.

For the **cohort selection** step, recent work has explored automatic methods based on rule driven systems (Vydiswaran et al., 2019) as well as NLP approaches that use discriminative language models (Soni and Roberts, 2021) or even generative large language models (Guan et al., 2023). Our method is related to rule based strategies but incorporates a hybrid module designed to improve generalizability, addressing a common limitation of traditional rule based cohort selection systems (Stubbs et al., 2019).

Regarding the **dataset generation** step, the quality of EHR data is a critical factor in ensuring the reliability and accuracy of predictive models, particularly in the context of medical research. Numerous studies have addressed the common challenges associated with EHR data quality, as well as methods for assessing and improving it (Cruz-Correia et al., 2009; Lewis et al., 2023; Feder, 2018; Terry et al., 2019). For instance, (Jetley and Zhang, 2019) suggest leveraging information extraction techniques or manual review of clinical notes to address gaps in structured data. This approach aligns with the hypothesis of the present project, which aims to utilize similar techniques for extracting relevant information from free-text discharge reports to supplement and improve the quality of structured EHR data.

The detrimental impact of missing values in clinical contexts is widely recognized. Numerous studies emphasize that missing data introduce uncertainty and bias into predictive models, presenting a significant challenge in this field (Sterne et al., 2009; Kahale et al., 2020; Ibrahim, Chu, and Chen, 2012). Consequently, implementing strategies to reduce the incidence of missing values can effectively mitigate the limitations of current predictive models.

The **labeling of instances** is a critical component of ED pipelines, yet producing high quality gold standard labels demands extensive expert involvement and manual review. This makes the process time consuming, costly, and difficult to scale. As a result, the use of automatically generated silver standard labels has become increasingly prevalent in ED research. Although these labels are imperfect, prior studies have shown that they can provide sufficient signal to train effective models (Wagholikar et al., 2020; McDavid et al., 2013).

In the particular context of **AF progression prediction**, research in this domain typically focuses on two primary scenarios: the prediction of new-onset AF and the recurrence of AF following therapeutic interventions. AI-driven models have demonstrated remarkable success in predicting incident AF, often outperforming conventional methods (Siontis et al., 2020). These models draw on diverse data sources such as clinical records, cardiac imaging, and electrophysiological data, including the use of EHRs (Tseng and Noseworthy, 2021; Nadarajah et al., 2021; Hulme et al., 2019; Tiwari et al., 2020). Notably, (Sung et al., 2022) developed a ML model to predict the risk of newly detected AF post-stroke, incorporating both structured variables and unstructured clini-

cal text processed through NLP.

Most existing studies on AF progression focus on the post-catheter ablation setting. As reviewed by (Fan et al., 2023), ML methods have shown strong performance in this context. These approaches often enhance predictive accuracy by combining clinical variables with additional echocardiography inputs (Knecht et al., 2024; Zhou et al., 2022), features extracted from electrocardiograms (ECGs) (Qiu et al., 2024), or anatomical data from computed tomography (Liu et al., 2024; Brahier et al., 2023).

In addition, despite all the advances in the field of AF progression prediction, a significant gap remains in accurately predicting AF progression following an initial event. This underexplored time window is clinically important, as it can support first-contact physicians with patient reference and cardiologists or arrhythmologists in making more informed treatment decisions. Consequently, our study systematically compares traditional clinical scores with our generated models using automatically obtained datasets.

3 Resources

The present research has been possible thanks to the following resources kindly provided by the Basque Public Healthcare System (Osakidetza):

- **Semi-structured discharge reports in Spanish**: 1.2×10^6 discharge reports from 2015 to 2020 (see Figure 2).

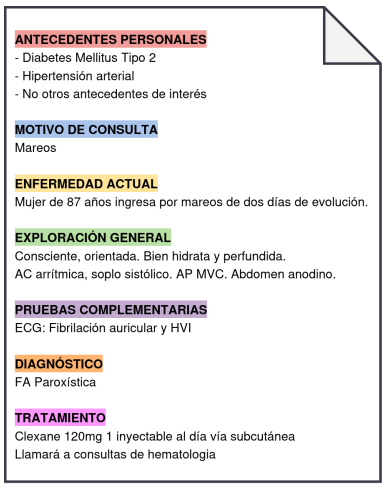

Figure 2: **Example of a semi-structured discharge report.**

- **Codified structured data**: Clinical data for each patient, encoded by healthcare professionals using standardized coding systems and stored in the Osakidetza Business Intelligence (OBI) platform (see Figure 3).

PatientId	Date	--	Test	Result
123342	11/8/2015	...	urea	33
...				
123548	22/1/2019	...	creatinine	0.71
123548	22/1/2019	...	PCR	3

Figure 3: **Example of the codified structured data.**

4 Experimental setup

The experimental setup focuses on the context of AF progression and is divided in three steps, the cohort selection for patients appropriate for the project, the tabular dataset generation step and the automatic labeling step of AF progression (See Figure 4).

4.1 Automatic Cohort selection

To study AF progression, we identified patients with AF onset, defined as their first documented AF episode with no prior history. Given the errors encountered in (Garcia Olea et al., 2021), we implemented a dual verification approach integrating both structured and unstructured data sources.

Patients were initially selected via structured EHR data (OBI system), then validated through clinical reports using a two-step NLP approach: first with a fine-tuned EriBERTa encoder-only language model (De la Iglesia et al., 2025), then a regular expression-based tool for quality control. Both tools are presented in (García-Olea et al., 2025).

4.2 Tabular Dataset Generation and Enrichment

A total of 85 clinical features were selected for this specific task, encompassing demographic data (7 features), patient history (35 features), laboratory results (18 features), procedures and their outcomes (7 features), treatments (16 features), and AF-related variables, including AF type and the AF progression status, which serves as the target label.

These features include established AF risk factors and general clinical markers aimed at facilitating the prediction of AF progression.

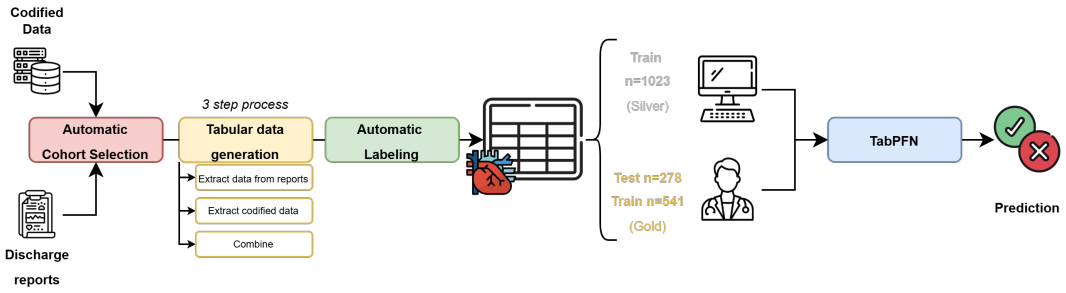

Figure 4: **End to end overview of the proposed methodology.** The pipeline starts with automatic cohort selection, followed by dataset generation by combining structured EHR data with information extracted from clinical reports. An NLP module performs automatic labeling, and the resulting silver and gold datasets are used to train and evaluate a TabPFN model for AF progression prediction.

While most of them are available and codified within the EHR, several key risk factors, such as left atrial size and even the AF progression status, are not represented in the structured coding system.

However, some of this information can be found in discharge reports, consequently it is important to retrieve this information to ensure the quality of the predictive models and to reduce the amount of manual annotation needed.

The tabular generation pipeline involved three key steps (see Figure 5):

1. **Extract and structure clinical information from discharge reports using the *Report2Vector* (R2V) pipeline.** This three-step NLP process detailed in (García-Olea et al., 2025) consists of: (a) section identification, which distinguishes between different parts of the report (e.g., past medical history vs. current episode) (de la Iglesia et al., 2023); (b) medical entity recognition, which extracts relevant clinical mentions such as symptoms, diagnoses, and procedures, along with negation detection to differentiate between confirmed and ruled-out conditions; and (c) regular expression matching to capture specific patterns of interest. This step returns a table with the 84 predictive clinical variables of interest for AF progression extracted from the discharge reports of the patient.

2. **Process structured EHR data from the OBI system using the *Structured2Vector* (S2V) module.** The same 84 predictive features obtained in the previous step are now extracted from the OBI system for each patient and organized into a tabular format.

3. **Merge both data sources into unified patient-level vectors using the *VectorMerger*.** The final tabular data contains the codified information from the OBI system enriched with the information extracted from the discharge reports.

4.3 Automatic Labeling

The automatic labeling process for determining AF progression status consists of the following steps:

1. Each patient report is processed following the same approach as the R2V tool (using a section identification module, a medical entity recognition and negation component, and regular expressions). This process extracts the AF status for each consultation date, categorizing it as AF episode, return to sinus rhythm, or no information. As a result, the complete history of the arrhythmia can be reconstructed for each patient.

2. Using the first documented AF episode as the onset point, subsequent consultations are examined for mentions of new AF episodes or returns to sinus rhythm. AF progression is defined as the occurrence of a new AF episode between one month and two years after the initial diagnosis. Based on this rule, three labels are assigned:

    - AF Progression (1): Explicit mention of new AF episode after its on-

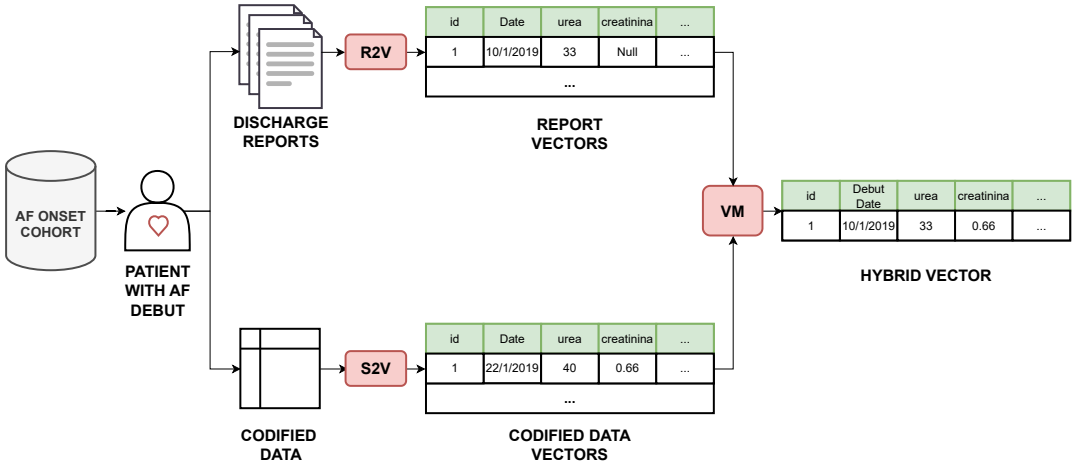

Figure 5: **Overview of the vector generation process.** For each patient in the AF onset cohort, all discharge reports (free text) and codified data (structured data stored in the Business Intelligence system) are collected and processed using the *Report2Vector (R2V)* and *Structured2Vector (S2V)* tools, respectively. Each tool generates a corresponding set of vectors, which are then merged by the *VectorMerger (VM)* tool to produce a patient-specific vector that integrates both sources of clinical information.

set.

- No Progression (0): Documented sinus rhythm or non-AF ECG findings after the onset.
- Excluded (-1): Cases without sufficient evidence to determine progression status.

4.4 Prediction of AF Progression

To evaluate the proposed dataset generation methodology, we also apply it to its original purpose: predicting atrial fibrillation (AF) progression within one month to two years after onset.

In this study, we employ a tabular foundation model, also referred to as a Large Tabular Model (LTM) (van Breugel and van der Schaar, 2024). Specifically, we use TabPFN (Hollmann et al., 2025) due to its ability to efficiently handle complex, small-scale tabular datasets while leveraging prior knowledge through pretraining.

During preliminary experimentation, several ML architectures were evaluated, including Support Vector Machines, Random Forests, and XGBoost. However, due to the high dimensionality of the predictive features and the substantial proportion of missing values, these models did not yield satisfactory results and were subsequently discarded. The experiments also explored various missing value imputation strategies (such as mean and median imputation, as well as logistic regression–based imputation) none of which outperformed the TabPFN performance. Additionally, data preprocessing techniques including feature standardization, feature selection, and different sampling strategies (undersampling, oversampling, SMOTE, and Tomek). Moreover, TabPFN incorporates internal data preprocessing strategies (including handling missing values and scale normalization) as well as an integrated hyperparameter optimization routine, which we utilized.

We evaluate performance differences between two datasets: the original dataset derived from codified EHR information, and the enriched dataset that incorporates features extracted from discharge reports.

In addition, we compare model performance when using automatically generated silver-standard annotations with that obtained using gold-standard labels manually annotated by a cardiologist. The characteristics of the dataset used in these experiments are summarized in Table 1.

4.4.1 Evaluation

Model performance was assessed using both accuracy and the Matthews Correlation Coefficient (MCC). While accuracy measures the proportion of correct predictions, it can be misleading in datasets with imbalanced classes. MCC provides a more robust evaluation by incorporating all elements of the con-

	Size	% Positives
Train-Silver	1023	65.40%
Train-Gold	541	66.17%
Test	278	64.03%

Table 1: **Characteristics of the dataset.** The second column indicates the number of patients and the last column the percentage of AF progression.

fusion matrix (true positives, true negatives, false positives, and false negatives) yielding values between –1 (complete disagreement) and 1 (perfect agreement).

In medical prediction tasks, MCC is especially valuable because it evaluates the quality of predictions for both classes simultaneously. Unlike the F1-score, which focuses solely on the positive class, MCC captures the balance between correctly identifying patients with and without the condition. This distinction is crucial in clinical contexts, where recognizing true negatives is as important as detecting true positives to prevent unnecessary interventions, costs, and patient distress. Consequently, MCC offers a more comprehensive and reliable measure of model performance in healthcare applications.

Moreover, to evaluate the clinical relevance of our predictive models, we compared them with established clinical scores for AF progression: CHADS2-VASc[2] (Lip et al., 2010), HATCH[3] (De Vos et al., 2010), and APPLE[4] (Kornej et al., 2015).

As these scores produce numerical values rather than direct classifications, a threshold of ≥ 2 was applied to convert them into binary outcomes. The scores were calculated using the generated tabular data and the formulas from their original publications.

5 Results and Discussion

In this section, we present the results of the proposed methodology, analyzing its impact on the proportion of missing values in the final dataset as well as the differences between the automatic and gold-standard annotations. We also report the outcomes of the AF progression prediction experiments performed using the different versions of the dataset.

[2]Used for stroke risk stratification in AF patients.

[3]Used for AF onset prediction.

[4]Used for AF recurrence risk prediction after catheter ablation.

5.1 Dataset Enrichment

The enriched dataset using the information of discharge reports improves considerably the amount of missing values and general recall of variables (see Figure 6). The comparison between the original and enriched datasets highlights the significant impact of incorporating information extracted from discharge reports.

As shown in the figure 6, the proportion of missing values in the original dataset (blue bars) is notably reduced in the enriched dataset (red bars) across most laboratory variables, indicating a substantial improvement in data completeness. Variables such as albumin, CRP, and NT-proBNP (which originally presented high levels of missingness) show a marked increase in data availability after enrichment. This suggests that the integration of textual information helps recover clinically relevant details often absent from coded records.

The recovery of past medical history features increased substantially, rising from an average of 2.62% positives in the original dataset to 14.25% positives in the enriched version (an absolute gain of 11.63 percentage points). Among all feature categories, treatment-related variables benefited the most from the enrichment process. Their mean recall rose from 1.73% in the original dataset to 45.05% in the enriched dataset.

Furthermore, certain clinically relevant variables with high predictive value were completely missing from the original codified dataset, as they are not recorded within the OBI system. For example, the left atrial size is a demonstrated predictor for AF recurrence and through the use of NLP extraction from discharge reports, this feature achieved a recall of 41.8%, which is remarkable given that not all patients have this information documented in their discharge summaries.

Missing values in EHRs can arise from multiple sources beyond human error. They may result from data integration issues across systems, variations in clinical documentation practices, or patient-related factors such as refusal or missed tests. System design limitations, such as non-mandatory fields also might contribute, as do temporal gaps when data is pending or historical records are unavailable. Overall, the enrichment process not only enhances data completeness but also improves the representativeness of clinical

variables, providing a stronger foundation for downstream predictive modeling.

5.2 Automatic Labeling

The automatic labeling approach for patients' AF progression status demonstrated an accuracy of 0.82 relative to the manually annotated test set. Achieving an accuracy of 0.82 is notable, especially considering that the manual annotations incorporate the full spectrum of patient information, including electrocardiograms and other details that may not be fully documented in the discharge reports. This means our methodology is able to capture and label patient AF progression status accurately despite the possible incompleteness of the discharge data.

5.3 AF Progression Prediction

To further assess the utility of our methodology and its application in future studies of early prediction of diseases we performed three experiments for AF progression prediction. The three experiments involve different versions of the generated dataset to evaluate how the proposed methodology impacts the results of AF progression. The first experiment uses the original dataset (1023 instances) extracted solely from codified EHRs with the silver annotation. The second experiment uses the enriched dataset (1023 instances), which combines codified information and discharge report data along with our silver-automatic annotation. Finally, the third experiment uses a smaller subset of the enriched dataset (541 instances) that has been manually annotated with gold-standard labels by a cardiologist.

As explained in section 4.4, we use the TabPFN architecture for the three experiments. The results are available in Table 2.

Dataset	Accuracy	MCC
Original-Silver	0.65	0.11
Enriched-Silver	0.66	0.20
Enriched-Gold	0.66	0.21

Table 2: **Results obtained in the AF progression experiments.**

Using the original dataset with silver annotations derived solely from codified EHRs, the model achieved a moderate accuracy of 0.65 and a low MCC of 0.11, indicating limited predictive power when relying exclusively on structured EHR data. Incorporating discharge report information in the enriched dataset with silver-automatic annotations led to an improvement in accuracy (0.66) and doubled the MCC (0.20), suggesting that integrating unstructured clinical text adds relevant information. In medical prediction tasks, this increase is particularly meaningful, as MCC reflects the overall reliability of the model across all possible outcomes. Doubling the MCC indicates a substantial improvement in the model's ability to correctly identify both patients at risk and those not at risk, which is crucial in clinical settings where misclassification can lead to missed diagnoses or unnecessary treatments.

Finally, using the manually annotated gold-standard subset resulted in a similar accuracy of 0.66, with a slight increase in MCC to 0.21. Although this subset contains almost half the number of instances, the higher-quality annotations improve the correlation between predictions and true labels. However, manual annotation is time-consuming and requires expert knowledge. Therefore, an automatic method that achieves comparable performance, even if it requires more instances, remains advantageous.

The predictive models consistently outperform the traditional clinical scores in both accuracy and MCC (see Table 3). Both CHADS2-VASc and HATCH achieved an accuracy of 0.60 with very low MCC values (–0.0052 and 0.0832, respectively). APPLE performed worse, with an accuracy of 0.48 and an MCC of 0.05.

	ACC	MCC
CHADS2-VASc	0.6043	-0.0052
HATCH	0.6043	0.0832
APPLE	0.4820	0.0510

Table 3: **Results obtained in the AF progression experiments by the clinical scores.**

These results highlight that data-driven predictive models, even with automatic annotations, can capture patterns in AF progression that traditional clinical scores fail to detect. While clinical scores are useful for quick risk stratification, their limited consideration of patient-specific information and simplified scoring rules result in lower predictive performance. In contrast, integrating structured EHR data with textual discharge reports al-

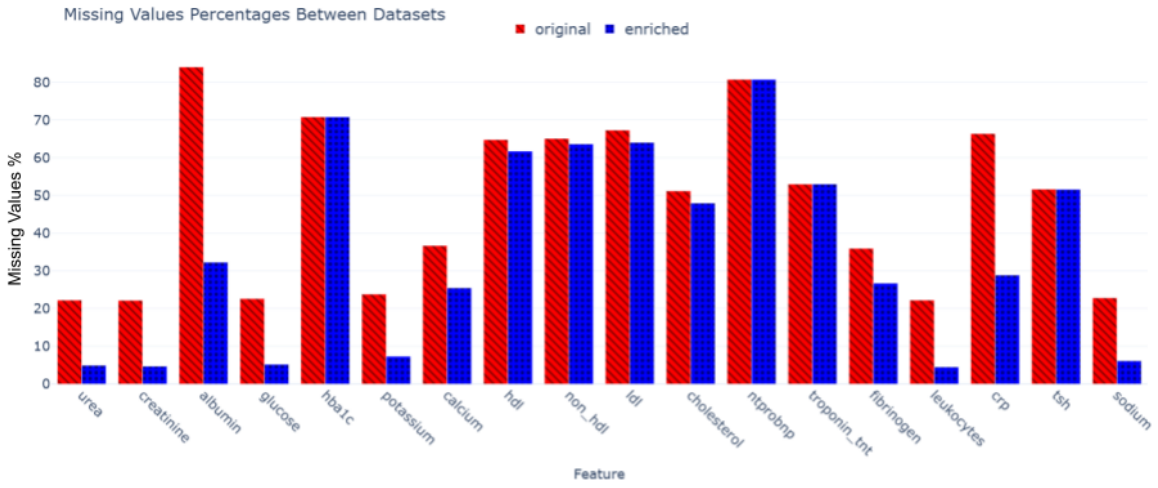

Figure 6: **Amount of missing values. Difference in percentage between the original and enriched datasets.** The bar plots illustrate the recovery of features that were absent in the original codified dataset but retrieved from the information contained in the discharge reports.

lows the models to leverage a richer set of features, yielding more reliable predictions, particularly for imbalanced outcomes like AF progression. This suggests that such models could serve as valuable decision-support tools, complementing or enhancing existing clinical scores.

6 Conclusions

This study provides evidence that discharge reports, when processed with NLP techniques, can play a significant role in supporting early prediction tasks and enhancing clinical data quality. Addressing each research question:

RQ1: Can the combination of structured EHR data and information extracted from discharge reports enhance and automate the development of early disease prediction models? The study demonstrates that integrating structured EHR data with NLP extracted information from discharge reports both facilitates automation of the data preparation pipeline and yields measurable improvements in model performance and data completeness. The proposed pipeline automates cohort selection, tabular feature generation and outcome labeling by combining codified EHR vectors with features extracted from free text. Practically, this automation reduces the need for manual case review and manual label assignment, streamlining the end to end process required to build early prediction models. In the AF progression experiments the enriched dataset that merged codified data with discharge report features produced a higher MCC than the codified dataset alone (MCC 0.20 versus 0.11) while accuracy rose slightly from 0.65 to 0.66. Automatic labeling via the pipeline also produced a high agreement with manual annotation, with an automatic labeling accuracy of 0.82 on the manually annotated test set. These results indicate that the integrated approach both automates previously manual steps and supplies additional predictive signal for the model.

RQ2: Can free-text discharge reports processed with NLP techniques improve the quality and completeness of structured tabular data derived from EHRs The integration of NLP-extracted information with existing tabular data substantially reduces missingness and increases recall for many clinically relevant variables that are underrepresented in codified fields. The article reports a clear reduction in the proportion of missing values for many laboratory and categorical features after enrichment with discharge report information. Laboratory variables that originally had high missingness, such as albumin, C reactive protein and NT proBNP, showed marked increases in availability. Similarly, recall for past medical history and treatment features rose noticeably in the enriched dataset. The VectorMerger approach produces uni-

fied patient vectors where features absent from the codified system are recovered from text, improving completeness and representativeness. This adresses the gaps caused by system limitations, documentation variability, or patient-related factors, the enrichment process enhances data completeness and representativeness, providing a more reliable foundation for predictive modelling.

RQ3: Can automatically generated silver annotations achieve performance comparable to gold-standard annotations while significantly reducing manual effort? Automatically generated silver annotations achieve performance close to gold-standard annotations in model evaluation and greatly reduce manual effort. The similarity of the enriched-silver and enriched-gold results, especially in MCC, indicates that silver annotations can approach gold performance in this problem setting. Given comparable model performance and that manual gold labeling is costly and time consuming, silver labeling permits larger training sets and faster iteration.

RQ4: Does the proposed methodology produce predictive models that outperform existing clinical scores for AF progression prediction? Models trained with the enriched dataset outperform traditional clinical scores for AF progression on both accuracy and MCC. Across the experiments, data driven models consistently outperformed the clinical scores considered. The TabPFN model trained on the enriched dataset achieved accuracy 0.66 and MCC 0.20. By contrast, CHADS2 VASc and HATCH each had accuracy approximately 0.60 with very low or near zero MCC values. The APPLE score performed worse with accuracy 0.48. These results indicate that ML models capture predictive patterns that simplified clinical scores do not.

Overall, this study highlights the potential of combining structured EHR data with NLP-extracted information from discharge reports to enhance early prediction tasks. Such an approach not only supports more robust and comprehensive early prediction models for AF progression but also lays the groundwork for applying similar methodologies to other diseases, ultimately contributing to more informed and data-driven clinical decision-making.

7 Future Work

The current study was limited to AF progression; applying the same pipeline to other chronic and acute conditions will be essential to assess its generalizability and robustness. In addition, external validation using datasets from different hospitals and healthcare systems will help evaluate the methodology under diverse documentation practices and data standards. Expanding the study to multiple institutions and involving a larger number of expert annotators would further strengthen the robustness and external validity of the proposed framework. However, generating high quality expert labels is a time consuming and resource intensive process, which was beyond the scope of the present work. Therefore, broader expert involvement and multi center validation are important directions for future research.

Future work will also explore the incorporation of advanced NLP approaches, including large language models (LLMs), into the existing pipeline. The rapid evolution of these architectures offers promising opportunities to further enhance the accuracy of feature extraction and automatic labeling.

In the specific context of AF progression prediction, future research could investigate multimodal approaches that process structured tabular data and unstructured clinical text jointly when estimating patient risk. Moreover, integrating temporal information from longitudinal patient histories may improve the modeling of disease dynamics and lead to more precise and clinically meaningful predictions.

Acknowledgements

This work has been partially supported by the HiTZ Center and the Basque Government, Spain (Research group funding IT1570-22) as well as by MCIN/AEI/10.13039/5011 00011033 Spanish Ministry of Universities, Science and Innovation by means of the projects: EDHIA PID2022-136522OB-C22 (also supported by FEDER, UE).

A. G. Domingo-Aldama has been funded by the Predoctoral Training Program for Non-PhD Research Personnel grant of the Basque Government (PRE_2024_1_0224).

A. García Olea has been funded by BioBizkaia grant under the code BB/I/PMIR/24/001.

References

Alzubi, A. A., V. J. Watzlaf, and P. Sheridan. 2021. Electronic health record (ehr) abstraction. *Perspectives in health information management*, 18(Spring).

Awwalu, J., A. G. Garba, A. Ghazvini, and R. Atuah. 2015. Artificial intelligence in personalized medicine application of ai algorithms in solving personalized medicine problems. *International Journal of Computer Theory and Engineering*, 7(6):439.

Botsis, T., G. Hartvigsen, F. Chen, and C. Weng. 2010. Secondary use of ehr: data quality issues and informatics opportunities. *Summit on translational bioinformatics*, 2010:1.

Brahier, M. S., F. Zou, M. Abdulkareem, S. Kochi, F. Migliarese, A. Thomaides, X. Ma, C. Wu, V. Sandfort, P. J. Bergquist, et al. 2023. Using machine learning to enhance prediction of atrial fibrillation recurrence after catheter ablation. *Journal of Arrhythmia*, 39(6):868–875.

Chen, H., X. Li, X. He, A. Chen, J. McGill, E. C. Webber, H. Xu, M. Liu, and J. Bian. 2025. Enhancing patient-trial matching with large language models: A scoping review of emerging applications and approaches. *JCO Clinical Cancer Informatics*, 9:e2500071.

Cruz-Correia, R. J., P. P. Rodrigues, A. Freitas, F. C. Almeida, R. Chen, and A. Costa-Pereira. 2009. Data quality and integration issues in electronic health records. In *Information discovery on electronic health records*. Chapman and Hall/CRC, pages 73–114.

De la Iglesia, I., A. Sánchez-Freire, O. Urquijo-Durán, A. Barrena, and A. Atutxa. 2025. Eriberta private surpasses her public alter ego: Enhancing a bilingual pretrained encoder with limited private medical data. *Procesamiento del Lenguaje Natural*, 75:283–296.

de la Iglesia, I., M. Vivó, P. Chocrón, G. de Maeztu, K. Gojenola, and A. Atutxa. 2023. An open source corpus and automatic tool for section identification in spanish health records. *J. Biomed. Informatics*, 145:104461.

De Vos, C. B., R. Pisters, R. Nieuwlaat, M. H. Prins, R. G. Tieleman, R.-J. S. Coelen, A. C. van den Heijkant, M. A. Allessie, and H. J. Crijns. 2010. Progression from paroxysmal to persistent atrial fibrillation: clinical correlates and prognosis. *Journal of the American College of Cardiology*, 55(8):725–731.

Fan, X., Y. Li, Q. He, M. Wang, X. Lan, K. Zhang, C. Ma, and H. Zhang. 2023. Predictive value of machine learning for recurrence of atrial fibrillation after catheter ablation: A systematic review and meta-analysis. *Reviews in Cardiovascular Medicine*, 24(11):315.

Feder, S. L. 2018. Data quality in electronic health records research: quality domains and assessment methods. *Western journal of nursing research*, 40(5):753–766.

Garcia Olea, A., J. Ormaetxe Merodio, A. Atutxa Salazar, I. Diez Gonzalez, I. Fernandez De La Prieta, M. Maeztu Rada, E. Amuriza De Luis, K. Ugedo Alzaga, U. Idiazabal Rodriguez, I. Pereiro Lili, et al. 2021. The role of congestive heart failure at atrial fibrillation onset in the data entry errors of electronic health records. In *EUROPEAN JOURNAL OF HEART FAILURE*, volume 23, pages 303–304. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA.

García-Olea, A., A. G. Domingo-Aldama, M. Merino, K. Gojenola, J. Goikoetxea, A. Atutxa, and J. M. Ormaetxe. 2025. The application of deep learning tools on medical reports to optimize the input of an atrial-fibrillation-recurrence predictive model. *Journal of Clinical Medicine*, 14(7):2297.

Guan, Z., Z. Wu, Z. Liu, D. Wu, H. Ren, Q. Li, X. Li, and N. Liu. 2023. Cohortgpt: An enhanced gpt for participant recruitment in clinical study.

Hollmann, N., S. Müller, L. Purucker, A. Krishnakumar, M. Körfer, S. B. Hoo, R. T. Schirrmeister, and F. Hutter. 2025. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326.

Holmes, J. H., J. Beinlich, M. R. Boland, K. H. Bowles, Y. Chen, T. S. Cook, G. Demiris, M. Draugelis, L. Fluharty,

P. E. Gabriel, et al. 2021. Why is the electronic health record so challenging for research and clinical care? *Methods of information in medicine*, 60(01/02):032–048.

Hulme, O. L., S. Khurshid, L.-C. Weng, C. D. Anderson, E. Y. Wang, J. M. Ashburner, D. Ko, D. D. McManus, E. J. Benjamin, P. T. Ellinor, et al. 2019. Development and validation of a prediction model for atrial fibrillation using electronic health records. *JACC: Clinical Electrophysiology*, 5(11):1331–1341.

Ibrahim, J. G., H. Chu, and M.-H. Chen. 2012. Missing data in clinical studies: issues and methods. *Journal of clinical oncology*, 30(26):3297–3303.

Jetley, G. and H. Zhang. 2019. Electronic health records in is research: Quality issues, essential thresholds and remedial actions. *Decision Support Systems*, 126:113137.

Jin, Q., Z. Wang, C. S. Floudas, F. Chen, C. Gong, D. Bracken-Clarke, E. Xue, Y. Yang, J. Sun, and Z. Lu. 2024. Matching patients to clinical trials with large language models. *Nature communications*, 15(1):9074.

Kahale, L. A., A. M. Khamis, B. Diab, Y. Chang, L. C. Lopes, A. Agarwal, L. Li, R. A. Mustafa, S. Koujanian, R. Waziry, et al. 2020. Potential impact of missing outcome data on treatment effects in systematic reviews: imputation study. *bmj*, 370.

Knecht, S., J. Cyriac, P. Badertscher, P. Krisai, V. Schlageter, S. Osswald, M. Zellweger, M. Kuhne, and C. Sticherling. 2024. Machine learning for outcome prediction of atrial fibrillation recurrence after catheter ablation. *Europace*, 26(Supplement_1):euae102–556.

Kornej, J., G. Hindricks, M. B. Shoemaker, D. Husser, A. Arya, P. Sommer, S. Rolf, P. Saavedra, A. Kanagasundram, S. Patrick Whalen, et al. 2015. The apple score: a novel and simple score for the prediction of rhythm outcomes after catheter ablation of atrial fibrillation. *Clinical Research in Cardiology*, 104:871–876.

Lewis, A. E., N. Weiskopf, Z. B. Abrams, R. Foraker, A. M. Lai, P. R. Payne, and A. Gupta. 2023. Electronic health record data quality assessment and tools: a systematic review. *Journal of the American Medical Informatics Association*, 30(10):1730–1740.

Lip, G. Y., R. Nieuwlaat, R. Pisters, D. A. Lane, and H. J. Crijns. 2010. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*, 137(2):263–272.

Liu, C.-M., W.-S. Chen, S.-L. Chang, Y.-C. Hsieh, Y.-H. Hsu, H.-X. Chang, Y.-J. Lin, L.-W. Lo, Y.-F. Hu, F.-P. Chung, et al. 2024. Use of artificial intelligence and i-score for prediction of recurrence before catheter ablation of atrial fibrillation. *International Journal of Cardiology*, 402:131851.

McDavid, A., P. K. Crane, K. M. Newton, D. R. Crosslin, W. McCormick, N. Weston, K. Ehrlich, E. Hart, R. Harrison, W. A. Kukull, et al. 2013. Enhancing the power of genetic association studies through the use of silver standard cases derived from electronic medical records. *PloS one*, 8(6):e63481.

Mishra, J. and S. Tarar. 2020. Chronic disease prediction using deep learning. In *Advances in Computing and Data Sciences: 4th International Conference, ICACDS 2020, Valletta, Malta, April 24–25, 2020, Revised Selected Papers 4*, pages 201–211. Springer.

Nadarajah, R., J. Wu, A. F. Frangi, D. Hogg, C. Cowan, and C. Gale. 2021. Predicting patient-level new-onset atrial fibrillation from population-based nationwide electronic health records: protocol of find-af for developing a precision medicine prediction model using artificial intelligence. *BMJ open*, 11(11):e052887.

Ng, K., S. R. Steinhubl, C. DeFilippi, S. Dey, and W. F. Stewart. 2016. Early detection of heart failure using electronic health records: practical implications for time before diagnosis, data diversity, data quantity, and data density. *Circulation: Cardiovascular Quality and Outcomes*, 9(6):649–658.

Qiu, Y., H. Guo, S. Wang, S. Yang, X. Peng, D. Xiayao, R. Chen, J. Yang, J. Liu, M. Li,

et al. 2024. Deep learning-based multimodal fusion of the surface ecg and clinical features in prediction of atrial fibrillation recurrence following catheter ablation. *BMC Medical Informatics and Decision Making*, 24(1):225.

Ristevski, B. and M. Chen. 2018. Big data analytics in medicine and healthcare. *Journal of integrative bioinformatics*, 15(3):20170030.

Schork, N. J. 2019. Artificial intelligence and personalized medicine. *Precision medicine in Cancer therapy*, pages 265–283.

Shah, V. 2018. Next-generation artificial intelligence for personalized medicine: Challenges and innovations. *INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY*, 2(2):1–15.

Sharma, D. K., M. Chatterjee, G. Kaur, and S. Vavilala. 2022. Deep learning applications for disease diagnosis. In *Deep learning for medical applications with unique data*. Elsevier, pages 31–51.

Siontis, K. C., X. Yao, J. P. Pirruccello, A. A. Philippakis, and P. A. Noseworthy. 2020. How will machine learning inform the clinical care of atrial fibrillation? *Circulation research*, 127(1):155–169.

Soni, S. and K. Roberts. 2021. Patient cohort retrieval using transformer language models. In *AMIA annual symposium proceedings*, volume 2020, page 1150.

Sterne, J. A., I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter. 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338.

Stubbs, A., M. Filannino, E. Soysal, S. Henry, and Ö. Uzuner. 2019. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *Journal of the American Medical Informatics Association*, 26(11):1163–1171.

Sung, S.-F., K.-L. Sung, R.-C. Pan, P.-J. Lee, and Y.-H. Hu. 2022. Automated risk assessment of newly detected atrial fibrillation poststroke from electronic health record data using machine learning and natural language processing. *Frontiers in Cardiovascular Medicine*, 9:941237.

Terry, A. L., M. Stewart, S. Cejic, J. N. Marshall, S. de Lusignan, B. M. Chesworth, V. Chevendra, H. Maddocks, J. Shadd, F. Burge, et al. 2019. A basic model for assessing primary health care electronic medical record data quality. *BMC medical informatics and decision making*, 19:1–11.

Tiwari, P., K. L. Colborn, D. E. Smith, F. Xing, D. Ghosh, and M. A. Rosenberg. 2020. Assessment of a machine learning model applied to harmonized electronic health record data for the prediction of incident atrial fibrillation. *JAMA network open*, 3(1):e1919396–e1919396.

Tseng, A. S. and P. A. Noseworthy. 2021. Prediction of atrial fibrillation using machine learning: a review. *Frontiers in Physiology*, 12:752317.

Ullah, M., A. Akbar, and G. G. Yannarelli. 2020. Applications of artificial intelligence in early detection of cancer, clinical diagnosis and personalized medicine.

van Breugel, B. and M. van der Schaar. 2024. Why tabular foundation models should be a research priority. *Proceedings of the 41st International Conference on Machine Learning*, 235:48976–48993, 21–27 Jul.

Vydiswaran, V. V., A. Strayhorn, X. Zhao, P. Robinson, M. Agarwal, E. Bagazinski, M. Essiet, B. E. Iott, H. Joo, P. Ko, et al. 2019. Hybrid bag of approaches to characterize selection criteria for cohort identification. *Journal of the American Medical Informatics Association*, 26(11):1172–1180.

Wagholikar, K. B., H. Estiri, M. Murphy, and S. N. Murphy. 2020. Polar labeling: silver standard algorithm for training disease classifiers. *Bioinformatics*, 36(10):3200–3206.

Xie, S., Z. Yu, and Z. Lv. 2021. Multi-disease prediction based on deep learning: a survey. *Computer Modeling in Engineering & Sciences*, 128(2):489–522.

Yu, Z., K. Wang, Z. Wan, S. Xie, and Z. Lv. 2023. Popular deep learning algorithms for disease prediction: a review. *Cluster Computing*, 26(2):1231–1251.

Zhou, X., K. Nakamura, N. Sahara, T. Takagi, Y. Toyoda, Y. Enomoto, H. Hara, M. Noro, K. Sugi, M. Moroi, et al. 2022. Deep learning-based recurrence prediction of atrial fibrillation after catheter ablation. *Circulation Journal*, 86(2):299–308.