Journal of
Clinical Medicine

MDPI

Systematic Review

Many Models, Little Adoption—What Accounts for Low Uptake of Machine Learning Models for Atrial Fibrillation Prediction and Detection?

Yuki Kawamura [1,*], Alireza Vafaei Sadr [2], Vida Abedi [2] and Ramin Zand [3,*]

[1] School of Clinical Medicine, University of Cambridge, Cambridge CB3 0SP, UK
[2] Department of Public Health Sciences, College of Medicine, The Pennsylvania State University, Hershey, PA 17033, USA; vabedi@pennstatehealth.psu.edu (V.A.)
[3] Department of Neurology, College of Medicine, The Pennsylvania State University, Hershey, PA 17033, USA
* Correspondence: yk402@cam.ac.uk (Y.K.); rzand@pennstatehealth.psu.edu (R.Z.)

Abstract: (1) **Background:** Atrial fibrillation (AF) is a major risk factor for stroke and is often underdiagnosed, despite being present in 13–26% of ischemic stroke patients. Recently, a significant number of machine learning (ML)-based models have been proposed for AF prediction and detection for primary and secondary stroke prevention. However, clinical translation of these technological innovations to close the AF care gap has been scant. Herein, we sought to systematically examine studies, employing ML models to predict incident AF in a population without prior AF or to detect paroxysmal AF in stroke cohorts to identify key reasons for the lack of translation into the clinical workflow. We conclude with a set of recommendations to improve the clinical translatability of ML-based models for AF. (2) **Methods**: MEDLINE, Embase, Web of Science, Clinicaltrials.gov, and ICTRP databases were searched for relevant articles from the inception of the databases up to September 2022 to identify peer-reviewed articles in English that used ML methods to predict incident AF or detect AF after stroke and reported adequate performance metrics. The search yielded 2815 articles, of which 16 studies using ML models to predict incident AF and three studies focusing on ML models to detect AF post-stroke were included. (3) **Conclusions**: This study highlights that (1) many models utilized only a limited subset of variables available from patients' health records; (2) only 37% of models were externally validated, and stratified analysis was often lacking; (3) 0% of models and 53% of datasets were explicitly made available, limiting reproducibility and transparency; and (4) data pre-processing did not include bias mitigation and sufficient details, leading to potential selection bias. Low generalizability, high false alarm rate, and lack of interpretability were identified as additional factors to be addressed before ML models can be widely deployed in the clinical care setting. Given these limitations, our recommendations to improve the uptake of ML models for better AF outcomes include improving generalizability, reducing potential systemic biases, and investing in external validation studies whilst developing a transparent modeling pipeline to ensure reproducibility.

Keywords: machine learning; atrial fibrillation; prevention; detection; stroke; neural networks; decision trees; artificial intelligence

Citation: Kawamura, Y.; Vafaei Sadr, A.; Abedi, V.; Zand, R. Many Models, Little Adoption—What Accounts for Low Uptake of Machine Learning Models for Atrial Fibrillation Prediction and Detection? *J. Clin. Med.* **2024**, *13*, 1313. https://doi.org/10.3390/jcm13051313

Academic Editor: Nathan Wong

Received: 16 January 2024
Revised: 19 February 2024
Accepted: 23 February 2024
Published: 26 February 2024

1. Introduction

1.1. Background

A prominent risk factor for stroke is atrial fibrillation (AF), which increases the incidence of stroke by between 2.6- and 4.5-fold depending on decade of life [1] and recurrence of stroke by around 2-fold [2]. Given that 13–26% of patients with ischemic stroke have AF [3], prediction and management of AF hold promise in addressing the disease burden of stroke. Yet, screening and diagnosis of AF are not straightforward, especially amongst asymptomatic patients. In fact, 13.1% of patients with AF in the United States are estimated

to be undiagnosed, with 56% of undiagnosed patients belonging to higher stroke risk groups with a $CHADS_2$ score of 2 and above [4].

AF monitoring is complicated by the fact that up to a third of patients do not experience symptoms [5], which reduces the pretest probability. In absence of any symptomatic cues, AF episodes can only be captured through continuous long-term monitoring given their unpredictable nature, which can generate copious data. In addition, although Holter monitors are frequently used for long-term monitoring of AF, patient compliance is often a limiting factor for longer monitoring durations [6], possibly due to their bulkiness and need for applied leads. These challenges can hinder diagnosis of AF and initiation of anticoagulants for thromboembolism prevention, which is problematic given that recent studies have shown that AF can be detected after longer monitoring in between 20 and 30% of patients diagnosed with cryptogenic strokes [7–9].

Big data and machine learning (ML) algorithms can improve AF prediction and diagnosis in terms of both accuracy and throughput. Digital health approaches could especially be effective in AF monitoring, since relevant measurements including heart rate and ECGs can be obtained using wearable devices which are readily available, noninvasive, and need not be replaced regularly in contrast to, for example, continuous glucose monitors. Indeed, the low uptake barrier for utilizing wearables for AF monitoring is illustrated by the fact that clinical trials investigating the performance of wearables such as the Apple Watch® [10] and Fitbit® [11] were able to recruit upwards of 400,000 volunteers. Vast amounts of data generated by such wearables, coupled with the ability of ML algorithms to model complex datasets, can enhance thromboembolic stroke prevention primarily in two ways. Screening a general population and predicting incident AF can allow high-risk patients to be monitored more carefully for the primary occurrence of embolic stroke. In addition, continuous electrocardiogram (ECG) monitoring of patients after ischemic stroke can identify patients for whom anticoagulants should be started to prevent stroke recurrence. Hence, improved prediction and detection capabilities conferred via ML models can improve primary and secondary stroke prevention.

Despite potential advantages, however, clinical uptake of ML-based models has been slow, and only six devices for artificial intelligence (AI)-based AF detection have been approved in the United States and Europe between 2015 and 2020 [12]. This study therefore aims to characterize the barriers to incorporating ML models into the clinical workflow of AF management. We performed a scoping review of studies proposing or evaluating ML models for AF prediction and detection to illustrate the current state of the art and analyzed the limitations of each of the studies, which might hinder clinical uptake. Given the observed limitations, we propose several recommendations for future studies that could promote the clinical translation of ML models for AF care.

1.2. Scope and Key Questions

We sought to address the following key questions (KQ):

(KQ1) In adult patients without a known history of stroke or AF or cardiovascular comorbidities, what are the performance statistics, data features and processing steps, and limitations of ML models in predicting incidence of AF?
(KQ2) In adult patients with a previous history of stroke, what are the performance statistics, data features and processing steps, and limitations of ML models for AF detection?
A PICOTS (population, interventions, comparators, outcomes, timing, and setting) table with details on the key questions is shown below (Table 1).

Table 1. PICOTS (population, interventions, comparators, outcomes, timing, and setting) table for Key Questions addressed in the scoping review.

	Key Question 1	Key Question 2
Population	Adult patients without a known history of stroke or atrial fibrillation or cardiovascular comorbidities	Adult patients with a previous history of stroke
Interventions	ML models to predict incidence of atrial fibrillation	ML models to detect atrial fibrillation
Comparators	None	None
Outcomes	• Predictive performance of models (AUC or 2 × 2 table) • Input data features and data processing steps used in model development • Limitations hindering model incorporation into clinical practice, including those listed by authors	• Detection performance of models (AUC or 2 × 2 table) • Input data features and data processing steps used in model development • Limitations hindering model incorporation into clinical practice, including those listed by authors
Timing	Any observational cohort study	Any observational cohort study
Setting	Any setting	Any setting

2. Methods

2.1. Search Strategy

The scoping review was performed and written in accordance with the Enhancing the QUAlity and Transparency of Health Research (EQUATOR) and Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; Table S1) guidelines. MEDLINE, Embase, Web of Science, Clinicaltrials.gov, and ICTRP databases were searched for articles proposing or evaluating ML models (including tree-based models but not logistic regression models) to (1) predict the incidence of atrial fibrillation in a population with no known history of stroke or atrial fibrillation or (2) to detect atrial fibrillation in a population with a previous history of stroke. Original research articles in English published in peer-reviewed journals were included. Conference abstracts were also included in the search but were removed as duplicates if the same model was evaluated by the same group in a subsequent peer-reviewed article (i.e., if they were preliminary results for a subsequent published article). Databases were searched for manuscripts containing variants of "atrial fibrillation", "stroke", and "machine learning/artificial intelligence", combined with "predict" or "detect" as appropriate. Details of search terms can be found in the Supplementary Information.

2.2. Eligibility Criteria

Observational cohort studies evaluating the accuracy of ML models predicting the incidence of atrial fibrillation in an adult population without a history of AF or detecting AF in a stroke population in any setting were included. We selected studies that provided an area under the receiver operating characteristics curve (AUC) or data sufficient to create a 2 × 2 table and used the diagnosis of AF using ECG or discharge codes as a gold standard. For studies regarding the detection of AF in a stroke population, only those that defined previous incidence of stroke based on the diagnosis of stroke by a neurologist or hospitalist or discharge codes were included.

Studies using or proposing new clinical risk scores, studies using mortality rather than diagnosis as the endpoint, studies performed on a disease population with underlying cardiovascular comorbidities (such as chronic kidney disease or diabetes) or structural heart disease, and studies not authored in English were excluded.

2.3. Data Collection

Data extraction was performed in duplicate. After the removal of irrelevant records, screening was performed by two authors (YK, RZ) to identify studies that met the inclusion and exclusion criteria. Disagreements were resolved through discussion involving

a third reviewer (VA). Full-text reviews were performed for selected studies to identify model architecture, input data and list of variables, study population, data pre-processing including addressing missingness and potential selection bias, validation, results including model performance metrics, follow-up period, and availability of models and codes for reproducing and transparency. Model performance was rounded to two significant figures.

3. Results

3.1. AI for Primary Stroke Prevention: Prediction of Atrial Fibrillation in the General Population

3.1.1. Search Results and Study Characteristics

The search yielded 2234 results. After removing duplicates, 1352 studies were screened based on their title and abstracts, which resulted in the selection of 129 studies for further screening. Of these studies, full texts for three studies were not retrieved. After reviewing the full texts of 126 studies, 16 studies [13–28] met the inclusion/exclusion criteria (Figure S1, Table S2). The 16 identified studies were conducted in seven different countries: Seven in the United States, three in Korea, two in the United Kingdom, and one in Japan, Lebanon, Germany, and Taiwan (Figure 1, Table S3). The publication dates spanned five years, ranging from 2017 to 2022. Most of the studies were conducted retrospectively (14/16), whereas two were conducted prospectively in the United States. The majority of studies (10/16) curated their data in a database accessible to approved investigators.

Figure 1. Cohort Characteristics of Selected Studies. (**A**) Number of selected studies performed in each country, (**B**) World map color-coded by study population size.

3.1.2. Machine Learning Models: Characteristics and Performance Metrics

Of the 16 studies, two studies were validation studies on the same neural network-based model incorporating clinical variables, whereas two studies were validation studies of a separate convolutional neural network model (ECG-AI) developed for AF detection but repurposed for prediction (Table S4). Hence, the search yielded 14 different prediction models. Half (7/14) of the models used tree-based models, whereas 11 of the models used neural network-based models (Figure 2A). Of the tree-based models, five were random forest models, with the rest being Adaboost and gradient-boosted trees (including light-GBM and XGBoost). Only four of the models directly took ECG traces as an input for a convolutional neural network (CNN) model, but two additional studies incorporated ECG parameters (such as R-R intervals) in the analysis (Figure 2B, Table S4). In some cases, probability outputs from a CNN model predicting AF were used as input into a Cox regression model together with clinical data. Other model architectures used included support vector machines. The performance of predictive models as reported by the authors ranged from an AUC of 0.69 to 0.96 for simple neural network models, 0.72 to 0.84 for CNN-based neural networks, and 0.75 to 0.99 for tree-based models (Figure 2C, Table S4). The follow-up period ranged from 3 months to 16 years (Figure 2D, Table S3).

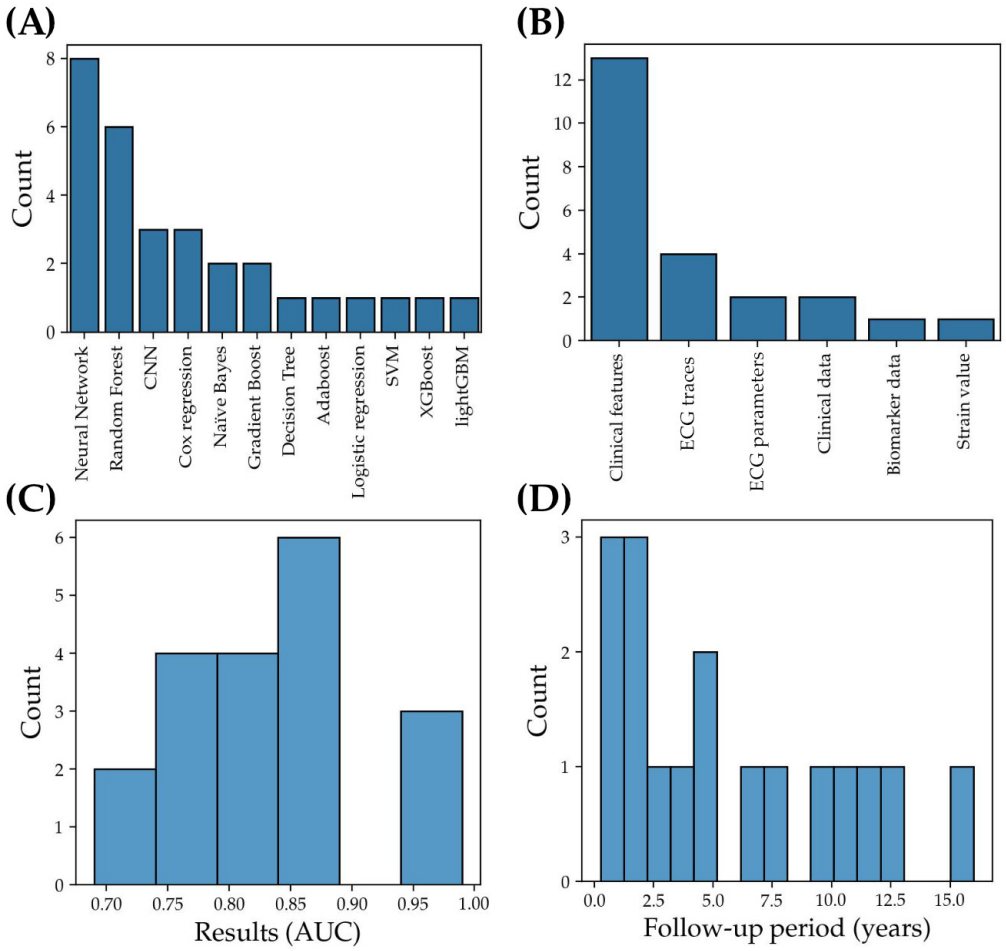

Figure 2. Model Characteristics of Selected Studies. (**A**) number of studies (count) for different model architectures, (**B**) number of studies (count) for different input data types, (**C**) number of studies (count) versus model AUC, and (**D**) number of studies (count) for different follow-up periods of selected studies.

3.1.3. Analysis of Limiting Factors and Best Practices: From Data Pre-Processing to Model Validation

Multiple factors limiting the real-world implementation of AI models for prediction and detection were identified (Figure 3, Table 2). One of the limiting factors was the lack of performance benchmarking against conventional predictive models, with the majority of studies (10/16) utilizing only one model architecture without comparing the performance of ML models against baseline models such as logistic regression (Figure 3A). The lack of external validation was another common shortfall, with only a subset (5/16) of studies reporting external validation and most of the studies relying on internal validation (Figure 3B). Furthermore, reporting of pre-processing to mitigate sparseness was limited, with only one study reporting adaptive imputation, two studies reporting complete case selection, and the other studies with no mention of pre-processing for sparseness (Figure 3C). In addition, the identified models often did not fully make use of the myriad data features that were available and instead used a limited subset of these features (Figure 3D, Table S5). Finally, none of the studies selected made their model or code available in public repositories (Table S4).

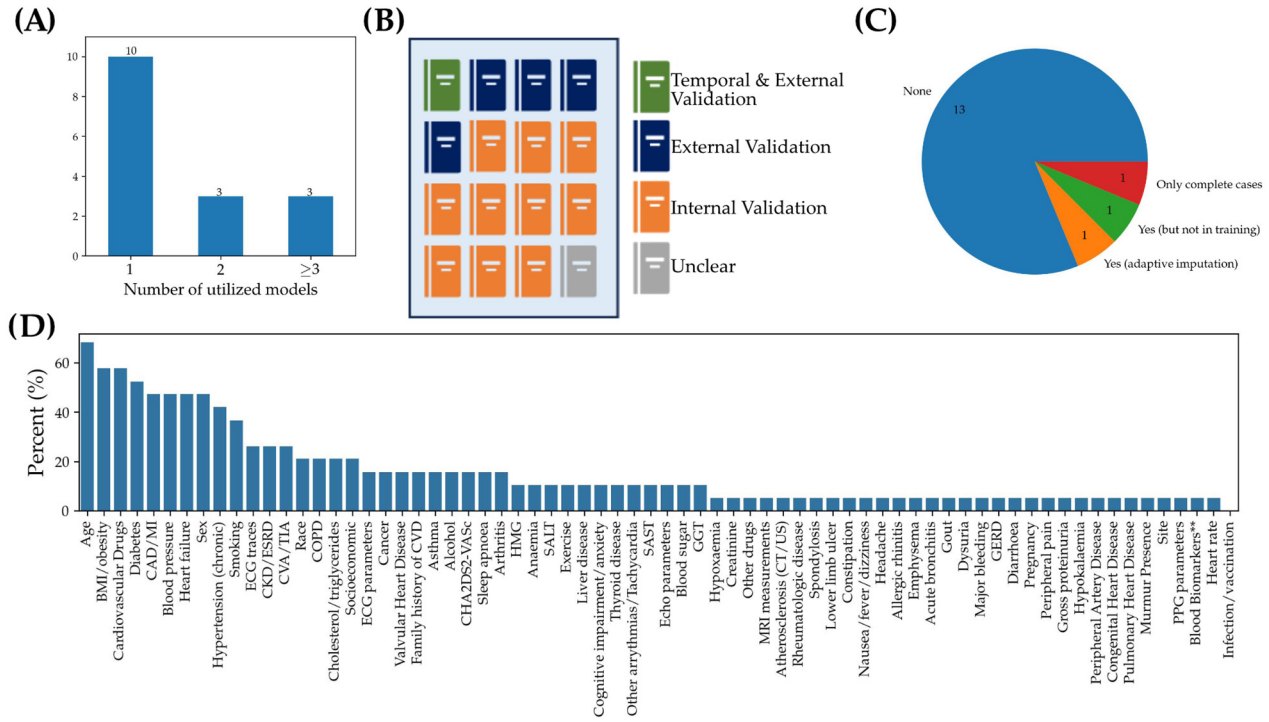

Figure 3. Data Processing in Selected Studies. (**A**) number of studies evaluating 1, 2, or more than 3 models, (**B**) type of validation performed in each of the 16 studies analyzed, (**C**) number of studies reporting sparse data processing, and (**D**) percent of studies for each input feature.

Table 2. Limitations of selected models for AF prediction.

Study (Original Study Proposing Model If Validation Study)	Limitations of the Study Suggested by the Authors	Additional Limitations
Ahmad et al., 2020 [13]	None listed	• Extremely small sample size • No external validation
Ambale-Venkatesh et al., 2017 [14]	• Patient cohort might not be representative of general population • Did not include genetic data • Longitudinal changes in risk not considered	• No external validation
ECG-AI (Attia et al., 2019 [29], Christopoulos et al., 2020 [15], Kaminski et al., 2022 [20])	• Lack of interpretability makes it difficult to direct therapy • Patient cohort (training and external validation) might not be representative of general population • Datasets could have been mislabeled	• Low PPV (3.2% at 95% sensitivity threshold) *
Hill et al., 2019 [16], Sekelj et al., 2021 [27]	• Improvement in accuracy was small • Patient cohort might not be representative of general population • No accounting for ethnic differences • Datasets could have been mislabeled • Potentially low cost-effectiveness	• Low PPV (12% at 75% sensitivity threshold)
Hirota et al., 2021 [17]	• Patient cohort might not be generalizable • ECG signals might not be generalizable to other devices	• No external validation
Hu et al., 2019 [18]	• Possibility of unaccounted confounding factors • Lacking information about lifestyle or family history of AF	• No external validation

Table 2. *Cont.*

Study (Original Study Proposing Model If Validation Study)	Limitations of the Study Suggested by the Authors	Additional Limitations
Joo et al., 2020 [19]	None listed	• No external validation • Relatively low AUC
Khurshid et al., 2022 [21]	• Potential selection bias • Limited sample size • Prediction window was too long/too short • Potential lack of interpretability	• Calibration of analysis demonstrates relatively mediocre performance of AI model in isolation (when not combined with CHARGE-AF)
Kim et al., 2020 [22]	• Potential selection bias • Limited precision regarding time of AF occurrence • Possible confounding variables • No external validation	
Kim et al., 2020 [23]	• Model only marginally outperformed clinical risk score when using same number of inputs	• No external validation
Lip et al., 2022 [24]	• Potential selection bias	• No external validation
Raghunath et al., 2021 [25]	• Limited AF monitoring could have led to missed AF occurrences. • Limited time of AF recording • Patient cohort might not be representative of general population • Lack of interpretability makes it difficult to find pathophysiological basis of prediction and establish causality	• No true external validation
Schnabel et al., 2023 [26]	• Patient cohort might not be representative of general population	• Low PPV (13% for 95% sensitivity threshold)
Tiwari et al., 2020 [28]	• Low sensitivity • Machine learning model did not significantly outperform logistic regression model and does not work in real time • No time-varying effects measured • Imprecise recording of when AF occurred • Did not incorporate biomarkers or laboratory values • Low PPV (5.9% at 75% sensitivity threshold) • Patient cohort might not be representative of general population	

PPV: Positive predictive value. * Low PPV is listed as a limitation only when their values are provided; this does not suggest that studies for which this limitation is not stated have higher PPVs.

3.2. AI for Secondary Stroke Prevention: Detection of Atrial Fibrillation in Stroke Cohorts

3.2.1. Search Results and Study Characteristics

The search yielded 581 articles. After removal of duplicates and further screening, three studies [30–32] were included in the review. Details on the study selection are summarized in Figure S2 and Table S6. Of the three identified studies, one was performed in Germany in 2014, one in the United States in 2019, and one in Taiwan in 2014 (Table S7). All the studies were performed prospectively on patients with ischemic stroke but differed slightly in their criteria, which are summarized in Table S7. None of the studies uploaded their data in a public repository.

3.2.2. Machine Learning Models: Characteristics and Performance Metrics

Of the three studies, two were validation studies on models developed previously, whereas one study proposed a new model (Table 3). The models were trained on R-R intervals in two of the studies, whereas one study used ECG traces directly (Table 3). The model architectures used were support vector machines (n = 2) and convolutional neural networks (n = 1). The sensitivity of detection models was between 63 and 95%, whereas the

specificity was between 35 and 96%. Reported positive predictive values ranged between 23 and 27%, whereas negative predictive values ranged between 94 and 96%.

Table 3. Characteristics of Models for Detection of Atrial Fibrillation in a Stroke Population.

Selected Study (Original Study Proposing Model If Validation Study)	Input Data	Data Source/Data Curated for Approved Access?	Model Architecture/Validation	Results	Model Interpretation	Code or Model Available/Reported Handling of Sparse Data	Model Currently Available for Clinical Use?
Rabinstein et al., 2021 [32] (ECG-AI [29])	ECG trace	Prospective; local EHR/no	CNN/External	Sn: 63% Sp: 75% PPV: 23% NPV: 94%	No	Neither/No	No
Reinke et al., 2018 [30]/ (Schaefer et al., 2014 [33])	ECG parameters	Prospective; local EHR/no	SVM (Proprietary model)/External	Sn: 95% Sp: 35% PPV: 27% NPV: 96%	No	Neither/No	Yes
Shan et al., 2014 [31]	Photoplethysmogram data	Prospective; local EHR/no	SVM/Internal	Acc: 96% Sn: 94% Sp: 96% AUC: 0.97	No	Neither/No	No

Sn: Sensitivity; Sp: Specificity; PPV: Positive predictive value; NPV: Negative predictive value; AUC: Area under the curve.

3.2.3. Analysis of Limiting Factors and Best Practices: From Data Pre-Processing to Model Validation

Limitations identified by the study authors included the limited size and setting of the cohort and limited monitoring duration (Table 4). Other limitations included low positive predictive value and low sensitivity. Furthermore, whilst two of the studies validated previously published models, one of the studies used a proprietary model whose development was not available. Common shortfalls regarding model training included the non-generalizability of the training cohort, as well the as lack of external validation. Other shortfalls concerned the quality of trained models, such as only marginal improvement on existing clinical risk scores and low positive predictive values (PPV). Additional limiting factors that could hinder real-life implementation included the lack of interpretability.

Table 4. Limitations of Selected Models for AF Detection.

Study	Limitations of the Study Suggested by the Authors	Additional Limitations
Rabinstein et al., 2021 [32]	• Small patient cohort • Short duration for AF monitoring • Reduced performance when adjusted for age	• Low sensitivity (63%) • Low PPV (23%) *
Reinke et al., 2018 [30]	• Monocentric design and small cohort • Proprietary algorithm which has not been thoroughly field-tested • Limited use of multimodal data	• Low PPV (27%) • No parameters/intermediate weights available (black-box model)
Shan et al., 2014 [31]	None listed	• No external validation

PPV: Positive predictive value. * Low PPV is listed as a limitation only when their values are provided; this does not suggest that studies for which this limitation is not stated have higher PPVs.

4. Discussion

4.1. Recommendations for Clinical Implementation of ML Models

Our study identified 19 studies proposing or validating ML-based models for the prediction and detection of AF. This review contributes insights into clinically relevant topics with limited prior attention, such as the reasons for low clinical translation of ML models as well as ML models for post-stroke AF detection. Whilst the results underscored the ability of ML models to outperform clinical risk scores and logistic regression, analyzing the studies revealed limitations in study design and model construction that negatively affected generalizability and thus could affect clinical translation. In particular, we identified four concerns which could most significantly hinder clinical uptake and provide recommendations for each below.

Firstly, many of the ML models did not include a full picture of the patients' health status and demographic information potentially available in electronic health records and

instead used a limited subset of data features. Furthermore, information regarding imaging, laboratory values, and blood biomarkers were not frequently included as input variables in the model. Including multimodal information as inputs in ML models could better aid clinicians in decision-making as it more closely mimics diagnostic reasoning and can also improve model performance. We suggest that future models take advantage of multimodal data available in generating a more holistic and clinically relevant decision.

Second, robust external validation should be performed to improve model generalizability. Given that complex ML models often overfit training data, robust external validation is essential in ensuring reliable results. However, only a small subset of selected studies included external validation, with the majority of studies relying on internal validation using data with the same demographic characteristics as the training set. Clinicians might be disinclined to adopt algorithms lacking external validation, since they might consider the evidence base to be less rigorous in comparison to clinical evidence based on consensus derived from multiple large trials. Decentralized training methods, such as federated learning and swarm learning, hold potential promise in surmounting privacy concerns whilst ensuring models are appropriately trained and tested on cohorts across multiple health systems [34,35]. We recommend that future studies incorporate external validation or use decentralized training methods to ensure that the ML model generates reliable results on unseen cohorts.

Thirdly, improving code, model, and data availability is essential to ensure transparency and promote clinicians' confidence in using ML-based models. Indeed, our analysis highlighted the disparity in cohort characteristics such as AF prevalence and ethnic composition as well as input variables and data formats amongst different studies, all of which can hinder generalizability. The availability of both the source code and data in a readily accessible and standardized fashion is necessary to promote models that can be deployed widely and effectively. Yet, none of the reviewed studies made their source code or even a black-box model publicly available online. Additionally, whilst the majority of studies on AF prediction curated their data in a dataset accessible to approved investigators, none of the studies for AF detection curated their data in such databases. Importantly, only one of the studies curated its data in a standardized dataset format, such as the Observational Medical Outcomes Partnership (OMOP) Common Dataset model [36], or the Patient-Centered Outcomes Research (PCORnet) Common Data Model (CDM) [37]. We recommend that publishing requirements strongly encourage the source code or the trained model to be made available, as well as de-identified data where appropriate.

Finally, care must be taken in data pre-processing and the deployment of ML models so that existing biases in healthcare are not unwittingly perpetuated. Indeed, a commercially available model in current use falsely assigned lower risk values to Black patients because it was trained using healthcare costs as a proxy for the severity of condition, despite the fact that on average less healthcare costs are spent on Black patients [38]. Clinical use of certain ML models might also require patients to have access to digital devices, which can be a hurdle for some high-risk populations including older patients and underprivileged patients [39], who risk being excluded from valuable studies. Possible biases were also identified in our analysis, such as the fact that most of the studies were performed in developed countries with high volumes of digital data, as well as selection bias for patients able to afford tertiary care, especially in the United States. Asian and Caucasian populations were well-represented, but no studies were performed in Africa or South America. We suggest maximizing diversity in training cohorts, incorporating fairness audits as standard practice, and post-processing to mitigate bias and ensure that ML models are amenable to use in a wide variety of contexts.

4.2. Further Considerations for ML Models in Clinical Practice

Effective incorporation of ML models in clinical practice requires not only the availability of high-performing and generalizable models but also support for both physicians and patients to reap maximal benefits. For example, the nature of tort law which privileges

standard of care regardless of its effectiveness in a particular case could put physicians at risk of liability should care be withheld based on risk assessment using ML models [40]. Implementing a clear legal framework on the use of and liability regarding ML models in clinical practice will aid physicians in defining how to incorporate model output in the clinical workflow and is likely to boost their uptake. Support for patients is just as important; a study comparing five commercially available wearable devices for AF monitoring demonstrated that approximately 20% of photoplethysmography or ECG traces were inconclusive due to artefact, which was a much higher rate than the rate published by the manufacturers [41]. Educating patients and promoting awareness of correct device use could be effective in improving the quality of collected data. Clinical trials validating the real-world effectiveness of ML-guided interventions would provide insights into ways in which human-AI interactions can be optimized.

Additionally, cost-effectiveness for the implementation of AF screening must also be considered. Given that AF prediction for a general population requires no invasive investigation and stroke patients with suspicion of cardioembolic stroke already receive telemetry, implementation of ML models itself might not be a significant economic burden. An analysis using the prediction model proposed by Hill et al., 2019 [16] suggested that AF screening will increase costs by approximately GBP 322 million but will also result an increase of 81,000 QALYs [42] in the UK population. However, screening a wide population, especially with a low pre-test probability of AF, could lead to an increased burden on the healthcare system due to more visits and confirmatory testing in a predominantly healthy, younger population. Indeed, the 2023 ACC/AHA/ACCP/HRS guidelines state that it has not been established that patients deemed to be of high risk of developing AF by validated risk scores benefit from screening and interventions [43]. Choosing the right target population and creating a clinical and legal framework which avoids the need for additional follow-up clinician visits are ways in which cost-effectiveness could be improved.

4.3. Limitations of the Study

This study has several limitations. First, only a limited number of studies were found regarding the detection of AF in stroke patients, because whilst there were many algorithms tested on a general population, there were few which were specifically tested on a stroke cohort. Future studies could externally validate existing algorithms on stroke cohorts to evaluate the performance of ML models on this clinically relevant population. Second, a quantitative comparison of different models was precluded because of significant differences in the types of models and variations in the input data, making it impossible to compare the performances of the models directly with one another. Studies comparing the performance of multiple models on an identical unseen dataset could provide insights into the generalizability of currently available ML models.

4.4. Conclusions

In conclusion, ML models hold promise in accurately predicting AF in a healthy population for primary prevention or detecting AF in a stroke population for secondary stroke prevention. Yet, their real-world deployment is currently limited. Our recommendations to improve clinical translation include improving generalizability, reducing potential systemic biases, and investing in external validation studies whilst developing a transparent modeling pipeline to ensure reproducibility. Future developments addressing these points would facilitate the much-needed translation of these models into the clinic.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jcm13051313/s1, Table S1. PRISMA checklist. Table S2. Excluded Studies for Prediction of Incident Atrial Fibrillation in Population without Prior AF. Table S3. Population Characteristics of Studies for Prediction of Atrial Fibrillation in General Population. Table S4. Characteristics of Models for Prediction of Atrial Fibrillation in the General Population. Table S5. Features Used in Final Model Training. Table S6. Excluded Studies for Detection of Atrial Fibrillation

after Stroke. Table S7. Population Characteristics of Studies for Detection of Atrial Fibrillation in a Stroke Population. Figure S1. PRISMA Flow Diagram of Studies Describing Prediction of Incident Atrial Fibrillation in a Population without Prior AF. Figure S2. PRISMA Flow Diagram of Studies Describing Detection of Atrial Fibrillation in a Stroke Population.

Author Contributions: Conceptualization, Y.K., V.A. and R.Z.; Methodology, Y.K.; Formal Analysis, Y.K. and R.Z.; Investigation, Y.K.; Data Curation, Y.K. and A.V.S.; Writing—Original Draft Preparation, Y.K.; Writing—Review and Editing, Y.K., A.V.S., V.A. and R.Z.; Visualization, A.V.S.; Supervision, V.A. and R.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wolf, P.A.; Abbott, R.D.; Kannel, W.B. Atrial fibrillation as an independent risk factor for stroke: The Framingham Study. *Stroke* **1991**, *22*, 983–988. [CrossRef]
2. Penado, S.; Cano, M.; Acha, O.; Hernández, J.L.; Riancho, J.A. Atrial fibrillation as a risk factor for stroke recurrence. *Am. J. Med.* **2003**, *114*, 206–210. [CrossRef]
3. Seiffge, D.J.; Werring, D.J.; Paciaroni, M.; Dawson, J.; Warach, S.; Milling, T.J.; Engelter, S.T.; Fischer, U.; Norrving, B. Timing of anticoagulation after recent ischaemic stroke in patients with atrial fibrillation. *Lancet Neurol.* **2019**, *18*, 117–126. [CrossRef]
4. Turakhia, M.P.; Shafrin, J.; Bognar, K.; Trocio, J.; Abdulsattar, Y.; Wiederkehr, D.; Goldman, D.P. Estimated prevalence of undiagnosed atrial fibrillation in the United States. *PLoS ONE* **2018**, *13*, e0195088. [CrossRef] [PubMed]
5. Healey, J.S.; Connolly, S.J.; Gold, M.R.; Israel, C.W.; Van Gelder, I.C.; Capucci, A.; Lau, C.P.; Fain, E.; Yang, S.; Bailleul, C.; et al. Subclinical Atrial Fibrillation and the Risk of Stroke. *N. Engl. J. Med.* **2012**, *366*, 120–129. [CrossRef]
6. Haowen, J.; Shyn Yi, T.; Jeremy King, W.; Jiaqi, L.; Tian Ming, T.; Vern Hsen, T.; Colin, Y. A meta-analysis of extended ECG monitoring in detection of atrial fibrillation in patients with cryptogenic stroke. *Open Heart* **2022**, *9*, e002081. [CrossRef]
7. Etgen, T.; Hochreiter, M.; Mundel, M.; Freudenberger, T. Insertable Cardiac Event Recorder in Detection of Atrial Fibrillation After Cryptogenic Stroke. *Stroke* **2013**, *44*, 2007–2009. [CrossRef] [PubMed]
8. Tayal, A.H.; Tian, M.; Kelly, K.M.; Jones, S.C.; Wright, D.G.; Singh, D.; Jarouse, J.; Brillman, J.; Murali, S.; Gupta, R. Atrial fibrillation detected by mobile cardiac outpatient telemetry in cryptogenic TIA or stroke. *Neurology* **2008**, *71*, 1696–1701. [CrossRef] [PubMed]
9. Ziegler, P.D.; Koehler, J.L.; Mehra, R. Comparison of continuous versus intermittent monitoring of atrial arrhythmias. *Heart Rhythm* **2006**, *3*, 1445–1452. [CrossRef]
10. Perez, M.V.; Mahaffey, K.W.; Hedlin, H.; Rumsfeld, J.S.; Garcia, A.; Ferris, T.; Balasubramanian, V.; Russo, A.M.; Rajmane, A.; Cheung, L.; et al. Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation. *N. Engl. J. Med.* **2019**, *381*, 1909–1917. [CrossRef]
11. Lubitz, S.A.; Faranesh, A.Z.; Selvaggi, C.; Atlas, S.J.; McManus, D.D.; Singer, D.E.; Pagoto, S.; McConnell, M.V.; Pantelopoulos, A.; Foulkes, A.S. Detection of Atrial Fibrillation in a Large Population Using Wearable Devices: The Fitbit Heart Study. *Circulation* **2022**, *146*, 1415–1424. [CrossRef] [PubMed]
12. Muehlematter, U.J.; Daniore, P.; Vokinger, K.N. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–2020): A comparative analysis. *Lancet Digit. Health* **2021**, *3*, e195–e203. [CrossRef] [PubMed]
13. Ahmad, A.; Mansour, S.; Zgheib, A.; Safatly, L.; Hajj, A.E.; Baydoun, M.; Ghaziri, H.; Aridi, H.; Ismaeel, H. Using Artificial Intelligence to Uncover Association of Left Atrial Strain with The Framingham Risk Score for Atrial Fibrillation Development. *J. Am. Coll. Cardiol.* **2020**, *75*, 455. [CrossRef]
14. Ambale-Venkatesh, B.; Yang, X.; Wu, C.O.; Liu, K.; Hundley, W.G.; McClelland, R.; Gomes, A.S.; Folsom, A.R.; Shea, S.; Guallar, E.; et al. Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. *Circ. Res.* **2017**, *121*, 1092–1101. [CrossRef]
15. Christopoulos, G.; Graff-Radford, J.; Lopez, C.L.; Yao, X.; Attia, Z.I.; Rabinstein, A.A.; Petersen, R.C.; Knopman, D.S.; Mielke, M.M.; Kremers, W.; et al. Artificial Intelligence-Electrocardiography to Predict Incident Atrial Fibrillation: A Population-Based Study. *Circulation. Arrhythmia Electrophysiol.* **2020**, *13*, e009355. [CrossRef]
16. Hill, N.R.; Ayoubkhani, D.; McEwan, P.; Sugrue, D.M.; Farooqui, U.; Lister, S.; Lumley, M.; Bakhai, A.; Cohen, A.T.; O'Neill, M.; et al. Predicting atrial fibrillation in primary care using machine learning. *PLoS ONE* **2019**, *14*, e0224582. [CrossRef]
17. Hirota, N.; Suzuki, S.; Arita, T.; Yagi, N.; Otsuka, T.; Kishi, M.; Semba, H.; Kano, H.; Matsuno, S.; Kato, Y.; et al. Prediction of current and new development of atrial fibrillation on electrocardiogram with sinus rhythm in patients without structural heart disease. *Int. J. Cardiol.* **2021**, *327*, 93–99. [CrossRef]

18. Hu, W.S.; Hsieh, M.H.; Lin, C.L. A novel atrial fibrillation prediction model for Chinese subjects: A nationwide cohort investigation of 682 237 study participants with random forest model. *Europace* **2019**, *21*, 1307–1312. [CrossRef]
19. Joo, G.; Song, Y.; Im, H.; Park, J. Clinical Implication of Machine Learning in Predicting the Occurrence of Cardiovascular Disease Using Big Data (Nationwide Cohort Data in Korea). *IEEE Access* **2020**, *8*, 157643–157653. [CrossRef]
20. Kaminski, A.E.; Albus, M.L.; Ball, C.T.; White, L.J.; Sheele, J.M.; Attia, Z.I.; Friedman, P.A.; Adedinsewo, D.A.; Noseworthy, P.A. Evaluating atrial fibrillation artificial intelligence for the ED: Statistical and clinical implications. *Am. J. Emerg. Med.* **2022**, *57*, 98–102. [CrossRef]
21. Khurshid, S.; Friedman, S.; Reeder, C.; Di Achille, P.; Diamant, N.; Singh, P.; Harrington, L.X.; Wang, X.; Al-Alusi, M.A.; Sarma, G.; et al. ECG-Based Deep Learning and Clinical Risk Factors to Predict Atrial Fibrillation. *Circulation* **2022**, *145*, 122–133. [CrossRef]
22. Kim, I.-S.; Yang, P.-S.; Jang, E.; Jung, H.; You, S.C.; Yu, H.T.; Kim, T.-H.; Uhm, J.-S.; Pak, H.-N.; Lee, M.-H.; et al. Long-term PM2.5 exposure and the clinical application of machine learning for predicting incident atrial fibrillation. *Sci. Rep.* **2020**, *10*, 16324. [CrossRef]
23. Kim, K.; Park, S.M. Artificial neural networks to compare the contribution of basic clinical factors, ESC SCORE, and multidimensional risk factors for cardiovascular event prediction performance: An observational study. *Eur. Heart J.* **2020**, *41*, 2897. [CrossRef]
24. Lip, G.Y.H.; Genaidy, A.; Tran, G.; Marroquin, P.; Estes, C. Incidence and Complications of Atrial Fibrillation in a Low Socioeconomic and High Disability United States (US) Population: A Combined Statistical and Machine Learning Approach. *Int. J. Clin. Pract.* **2022**, *2022*, 8649050. [CrossRef]
25. Raghunath, S.; Pfeifer, J.M.; Ulloa-Cerna, A.E.; Nemani, A.; Carbonati, T.; Jing, L.; vanMaanen, D.P.; Hartzel, D.N.; Ruhl, J.A.; Lagerman, B.F.; et al. Deep Neural Networks Can Predict New-Onset Atrial Fibrillation From the 12-Lead ECG and Help Identify Those at Risk of Atrial Fibrillation-Related Stroke. *Circulation* **2021**, *143*, 1287–1298. [CrossRef]
26. Schnabel, R.B.; Witt, H.; Walker, J.; Ludwig, M.; Geelhoed, B.; Kossack, N.; Schild, M.; Miller, R.; Kirchhof, P. Machine learning-based identification of risk-factor signatures for undiagnosed atrial fibrillation in primary prevention and post-stroke in clinical practice. *Eur. Heart J. Qual. Care Clin. Outcomes* **2022**, *9*, 16–23. [CrossRef] [PubMed]
27. Sekelj, S.; Sandler, B.; Johnston, E.; Pollock, K.G.; Hill, N.R.; Gordon, J.; Tsang, C.; Khan, S.; Ng, F.S.; Farooqui, U. Detecting undiagnosed atrial fibrillation in UK primary care: Validation of a machine learning prediction algorithm in a retrospective cohort study. *Eur. J. Prev. Cardiol.* **2021**, *28*, 598–605. [CrossRef]
28. Tiwari, P.; Colborn, K.L.; Smith, D.E.; Xing, F.; Ghosh, D.; Rosenberg, M.A. Assessment of a Machine Learning Model Applied to Harmonized Electronic Health Record Data for the Prediction of Incident Atrial Fibrillation. *JAMA Netw. Open* **2020**, *3*, e1919396. [CrossRef]
29. Attia, Z.I.; Noseworthy, P.A.; Lopez-Jimenez, F.; Asirvatham, S.J.; Deshmukh, A.J.; Gersh, B.J.; Carter, R.E.; Yao, X.; Rabinstein, A.A.; Erickson, B.J.; et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction. *Lancet* **2019**, *394*, 861–867. [CrossRef] [PubMed]
30. Reinke, F.; Bettin, M.; Ross, L.S.; Kochhäuser, S.; Kleffner, I.; Ritter, M.; Minnerup, J.; Dechering, D.; Eckardt, L.; Dittrich, R. Refinement of detecting atrial fibrillation in stroke patients: Results from the TRACK-AF Study. *Eur. J. Neurol.* **2018**, *25*, 631–636. [CrossRef]
31. Shan, S.M.; Tang, S.C.; Huang, P.W.; Lin, Y.M.; Huang, W.H.; Lai, D.M.; Wu, A.Y. Reliable PPG-based Algorithm in Atrial Fibrillation Detection. In Proceedings of the 2016 IEEE Biomedical Circuits and Systems Conference (BioCAS), Shanghai, China, 17–19 October 2016; pp. 340–343.
32. Rabinstein, A.A.; Yost, M.D.; Faust, L.; Kashou, A.H.; Latif, O.S.; Graff-Radford, J.; Attia, I.Z.; Yao, X.; Noseworthy, P.A.; Friedman, P.A. Artificial Intelligence-Enabled ECG to Identify Silent Atrial Fibrillation in Embolic Stroke of Unknown Source. *J. Stroke Cerebrovasc. Dis. Off. J. Natl. Stroke Assoc.* **2021**, *30*, 105998. [CrossRef] [PubMed]
33. Schaefer, J.R.; Leussler, D.; Rosin, L.; Pittrow, D.; Hepp, T. Improved Detection of Paroxysmal Atrial Fibrillation Utilizing a Software-Assisted Electrocardiogram Approach. *PLoS ONE* **2014**, *9*, e89328. [CrossRef] [PubMed]
34. Konečný, J.; Brendan McMahan, H.; Yu, F.X.; Richtárik, P.; Theertha Suresh, A.; Bacon, D. Federated Learning: Strategies for Improving Communication Efficiency. *arXiv* **2016**, arXiv:1610.05492. [CrossRef]
35. Warnat-Herresthal, S.; Schultze, H.; Shastry, K.L.; Manamohan, S.; Mukherjee, S.; Garg, V.; Sarveswara, R.; Händler, K.; Pickkers, P.; Aziz, N.A.; et al. Swarm Learning for decentralized and confidential clinical machine learning. *Nature* **2021**, *594*, 265–270. [CrossRef] [PubMed]
36. Voss, E.A.; Makadia, R.; Matcho, A.; Ma, Q.; Knoll, C.; Schuemie, M.; DeFalco, F.J.; Londhe, A.; Zhu, V.; Ryan, P.B. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J. Am. Med. Inform. Assoc.* **2015**, *22*, 553–564. [CrossRef] [PubMed]
37. Forrest, C.B.; McTigue, K.M.; Hernandez, A.F.; Cohen, L.W.; Cruz, H.; Haynes, K.; Kaushal, R.; Kho, A.N.; Marsolo, K.A.; Nair, V.P.; et al. PCORnet®2020: Current state, accomplishments, and future directions. *J. Clin. Epidemiol.* **2021**, *129*, 60–67. [CrossRef] [PubMed]
38. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **2019**, *366*, 447–453. [CrossRef]

39. Brandes, A.; Stavrakis, S.; Freedman, B.; Antoniou, S.; Boriani, G.; Camm, A.J.; Chow, C.K.; Ding, E.; Engdahl, J.; Gibson, M.M.; et al. Consumer-Led Screening for Atrial Fibrillation: Frontier Review of the AF-SCREEN International Collaboration. *Circulation* **2022**, *146*, 1461–1474. [CrossRef]
40. Price, W.N., II; Gerke, S.; Cohen, I.G. Potential Liability for Physicians Using Artificial Intelligence. *JAMA* **2019**, *322*, 1765–1766. [CrossRef]
41. Mannhart, D.; Lischer, M.; Knecht, S.; du Fay de Lavallaz, J.; Strebel, I.; Serban, T.; Vögeli, D.; Schaer, B.; Osswald, S.; Mueller, C.; et al. Clinical Validation of 5 Direct-to-Consumer Wearable Smart Devices to Detect Atrial Fibrillation: BASEL Wearable Study. *JACC Clin. Electrophysiol.* **2023**, *9*, 232–242. [CrossRef]
42. Hill, N.R.; Groves, L.; Dickerson, C.; Boyce, R.; Lawton, S.; Hurst, M.; Pollock, K.G.; Sugrue, D.M.; Lister, S.; Arden, C.; et al. Identification of undiagnosed atrial fibrillation using a machine learning risk prediction algorithm and diagnostic testing (PULsE-AI) in primary care: Cost-effectiveness of a screening strategy evaluated in a randomized controlled trial in England. *J. Med. Econ.* **2022**, *25*, 974–983. [CrossRef] [PubMed]
43. Joglar, J.A.; Chung, M.K.; Armbruster, A.L.; Benjamin, E.J.; Chyou, J.Y.; Cronin, E.M.; Deswal, A.; Eckhardt, L.L.; Goldberger, Z.D.; Gopinathannair, R.; et al. 2023 ACC/AHA/ACCP/HRS Guideline for the Diagnosis and Management of Atrial Fibrillation: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation* **2024**, *149*, e1–e156. [CrossRef] [PubMed]