Machine learning reveals hidden diagnoses among underserved patients

Divya Shanmugam[1,*], Brianna Hardy[2], Annabel Wang[3], Sanjay Divakaran[4], John Guttag[5], Emma Pierson[6,†], Michael Barnett[2,3,4,†]

[1]Cornell Tech, New York, NY, USA. [2]Chan School of Public Health, Cambridge, MA, USA. [3]Harvard Medical School, MA, USA. [4]Mass General Brigham, MA, USA. [5]Massachusetts Institute of Technology, Cambridge, MA, USA. [6]University of California, Berkeley, Berkeley, CA, USA. *Corresponding author. †Denotes co-senior-authorship.

Abstract

Diagnostic data in health records plays a central role in clinical practice and public health, but diagnoses are often incomplete. These "hidden diagnoses" may be more common for certain patient populations, obscuring disease burden and introducing bias into downstream research and health care delivery. In this study, we develop a new approach to measure hidden diagnoses and assess the potential magnitude of this bias. We apply it to atrial fibrillation (AF), a common, high-risk condition that is frequently undiagnosed. The approach consists of two steps. First, we use machine learning to estimate a patient's risk of AF from an initial, normal ECG. We then measure disparities in recorded diagnosis controlling for this estimated risk, allowing our approach to distinguish differences in diagnostic behavior from differences in disease risk. We find that Black and Hispanic patients are 18–32% less likely to be diagnosed with AF than white patients at the same estimated risk. Our estimates suggest that we would observe 50% more AF diagnoses among minorities if they had the same diagnosis rate as white patients. Our method overcomes limitations of prior approaches that rely on resource-intensive population screening or incomplete diagnostic data, and produces the first estimates of disparities in AF prevalence corrected for hidden diagnosis. Our findings highlight the potential of machine learning methods to uncover hidden inequities in care.

1 Introduction

Diagnoses recorded in electronic health records, via ICD-10 codes or other structured fields, are widely recognized as imperfect measures of disease. Patients with a disease can easily lack accurate labels due to substantial fragmentation in care delivery across providers, health systems, and electronic health records, among myriad other reasons. Despite their flaws, recorded diagnoses remain the standard for determining disease presence for both clinicians and researchers. Clinicians rely on diagnosis codes to guide treatment decisions and determine eligibility for therapies [1, 2]; health systems use them to evaluate quality of care [3, 4], track population-level disease burden [5, 6, 7], and recruit patients for trials [8, 9]; and algorithms increasingly use them as both inputs and outputs for predictive systems [10, 11]. Understanding the extent to which diagnoses fail to capture patients with disease is essential to ensuring the validity of their downstream usage.

These "hidden diagnoses" – cases where a patient has a disease but no label appears in the patient's electronic health record (EHR) – occur even when a disease is simple to diagnose. For example, numerous works have established high rates of hidden diagnosis in the context of chronic kidney disease, despite standard lab tests confirming disease presence [12, 13, 14, 15]. One cause of hidden diagnosis is *underdiagnosis*, where a condition is not recognized, but there are other causes as well. Importantly, hidden diagnoses are not evenly distributed: longstanding disparities in healthcare access [16, 17, 18, 19, 20] mean that certain patient groups are more likely to have hidden diagnoses compared to others. Such differences in diagnosis rates are consequential because they can, for example, distort estimates of disease prevalence, or introduce bias into machine learning models trained to predict diagnosis.

Here, we propose a novel approach to measuring disparities in hidden diagnosis: we train a machine learning model to estimate a patient's true risk of having a condition and measure differences in the probability of diagnosis conditional on estimated risk. This procedure allows us to quantify differences in hidden diagnosis across salient demographic groups including race/ethnicity, primary spoken language, and insurance coverage, following related past work in non-medical domains [21]. We study hidden diagnosis in the context of atrial fibrillation (AF), a common, widely underdiagnosed condition with life-threatening complications and multiple evidence-based therapies [22, 23, 24]. Specifically, for a patient with a normal ECG and no established AF, we use a deep neural network to estimate the patient's risk of atrial fibrillation (AF) by training the network to discriminate between (1) patients who subsequently experience an ECG in AF and (2) patients who do not. Current AF can be unambiguously detected from an ECG, thus providing a source of ground truth distinct from clinician-recorded diagnoses to train a neural network (Figure 1).

Prior efforts to measure hidden diagnosis of AF have provided valuable initial estimates, but face important limitations. Past work estimates hidden diagnosis by comparing *observed* diagnosis rates to *expected* diagnosis rates derived by either (1) screening a random sample of patients for disease presence [25, 26, 27, 28, 29, 30], or (2) inferring disease prevalence based on observed complications [31, 32, 33]. Screening-based approaches, while informative, require routine monitoring and are resource intensive, often struggling to recruit diverse patient populations. Complication-based approaches, meanwhile, inherit biases in data on recorded complications, where data quality can vary substantially across patient groups. Our approach addresses these limitations by estimating disease risk from a widely available physiological signal, enabling the measurement of hidden diagnosis across a much larger and more diverse patient population than prior work and allowing direct quantification of potential biases in recorded diagnosis.

We find that racial minorities experience a 18-32% lower rate of diagnosis relative to white patients with the same estimated risk. Our model additionally yields estimates for disparities in prevalence of AF between groups, and reveals that diagnosis-based methods to estimate AF prevalence substantially underestimate how common the condition is among underserved patient groups. We verify that our findings hold under a series of robustness checks. In sum, our results indicate that hidden diagnoses of AF are widespread, with approximately 50% of patients with established AF lacking a diagnosis, and disproportionately common among underserved patients. Our results help explain the longstanding "AF

paradox" —— the lower recorded prevalence of AF among Black patients despite their higher burden of risk factors compared to white patients [34, 35] —— by showing that a substantial portion of this gap stems from hidden diagnoses.

2 Results

Our goal is to assess the extent to which diagnoses of atrial fibrillation are "hidden" across multiple demographic groups. We first provide a conceptual overview of our approach and then describe how we implement it. The assumption underlying our approach is that patients at higher risk of AF should also be likelier to receive a diagnosis of AF. Thus, the first step in our method is to estimate a patient's risk of AF; to do so, we use machine learning to predict the probability that a patient who has a sinus rhythm ECG (i.e. an ECG with no rhythmic irregularities) will have an AF ECG within 90 days. The second step is to estimate demographic disparities in diagnosis rates *controlling* for risk. Patients with the same disease risk should receive diagnoses at similar rates; if this is not the case, it implies that some demographic groups are more likely to have hidden diagnoses than others when controlling for disease risk.

We implement the approach as follows. To estimate each patient's risk (step 1 of our approach), we train a deep learning model to predict a patient's risk of AF using a dataset of sinus rhythm ECGs collected between 2016 and 2019 in a large academic health system in the US. For clarity, we will refer to an AF identification via ECG as an *AF ECG* and an AF identification via EHR diagnosis as an *AF diagnosis*. Patients who go on to have an AF ECG within 90 days of their sinus rhythm ECG (n=9,518) serve as positive examples for our model to learn from; all remaining patients (n=396,517) serve as negatives. Model predictions thus capture a patient's probability of experiencing an AF ECG within 90 days. The final dataset contains 406,035 unique patients, divided into a training (n=243,824), calibration (n=101,381), and study (n=60,830) split, where no patient appears in any two splits. To estimate disparities conditional on risk (step 2 of our approach), we use the study split to fit a linear regression to estimate the presence of an *AF diagnosis* from the ECG-based estimate of AF risk, age, sex, healthcare utilization, hospital site, and one of the three groups we study (race, insurance, or primary language). Results for the joint regression (including all three demographic groups) are in Fig. 10. Coefficients of the linear regression associated with the demographic groups capture differences in diagnosis rates not explained by differences in estimated risk.

Preliminary evidence. The most basic test for hidden diagnosis is to examine which patients with an AF ECG (a ground truth indicator for AF) lack a diagnosis label in the electronic health record (EHR). We find that 51.4% of patients with ECG-confirmed AF (n=9,518) had an established diagnosis that was hidden because it lacked a structured label in the EHR (Table 1). Some patients receiving ECGs may not receive follow-up care in the same health system for a label to be documented (e.g. they have primary care in another state), so we also examined diagnosis labels for patients with an AF ECG who also have at least one PCP visit in the health system in the prior year (n=2,459, 25.8%). Even among this group, 1,026 (41.7%) did not have an EHR label for AF. Importantly, there were marked disparities in the rate of hidden diagnoses by patient group. Taking the subset with a prior PCP visit in the health system, 39.6% of white patients had a hidden diagnosis compared to 54.5% for Black patients. Hidden diagnosis rates are also higher among Medicaid patients (61.9%) compared to those who are commercially insured (40.6%), and higher among non-English speakers (48.5%) relative to English speakers (40.9%). Table 1 reports all results.

However, differences in hidden diagnosis rates among those with an AF ECG, as documented above, provide an imperfect assessment of disparities. In particular, the analysis is restricted to a small, non-representative sample (approximately 2.4% of the entire sample), limiting the statistical power and generalizability of the findings. Using machine learning to estimate AF risk from a sinus rhythm ECG, a widely

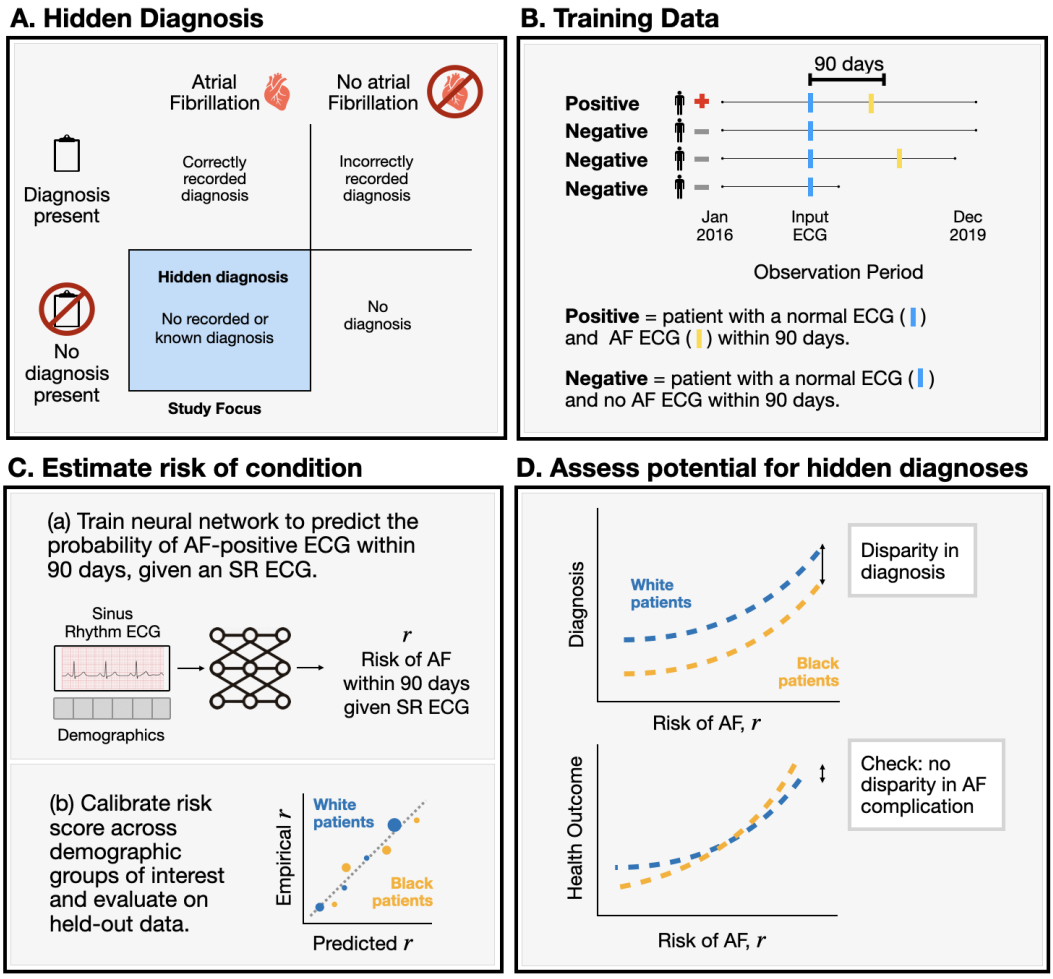

Figure 1: **Methodological overview.** Our goal is to measure the extent to which diagnosis are hidden: when a patient has atrial fibrillation (AF), but no AF diagnosis is present (Panel A). We construct training data using a dataset that captures ECGs for close to 500,000 patients over four years (Panel B). We define a positive example as a patient with a sinus rhythm (SR) ECG in the 90 days prior to an ECG in atrial fibrillation (AF) (ensuring no prior diagnosis of AF and complete demographic data, n=9,518). We define a patient with no observed AF ECGs, at least one sinus rhythm ECG, no prior diagnosis of AF, and complete demographic data to be negative (n=396,517). The first step of our approach (Panel C) is to estimate a patient's risk of an AF ECG within 90 days by training and calibrating a deep neural network on these data. The second step of our approach (Panel D) is to assess the potential for missed diagnosis by measuring differences in rates of *diagnosis* conditional on AF risk. We use disparities in *health* conditional on estimated risk as a robustness check.

		% **Hidden Diagnosis (CI)** among patients with AF ECG	% **Hidden Diagnosis (CI)** among patients with AF ECG, PCP Visit
Overall	–	51.4 (50.4, 52.4)	41.7 (39.8, 43.7)
Race	White	49.8 (48.7, 50.9)	39.6 (37.5, 41.8)
	Black/African American	56.8 (52.3, 61.1)	54.5 (46.9, 61.9)
	Hispanic/Latino	58.3 (52.8, 63.6)	50.0 (41.2, 58.8)
	Asian	56.8 (51.0, 62.5)	42.9 (31.9, 54.5)
Insurance	Commercial	53.4 (51.5, 55.3)	40.6 (36.5, 44.9)
	Medicare	46.3 (44.9, 47.7)	37.1 (34.5, 39.7)
	Medicaid	66.2 (62.1, 70.0)	61.9 (53.5, 69.7)
Primary Language	English	50.8 (49.8, 51.9)	40.9 (38.8, 42.9)
	Non-English	56.7 (53.6, 59.8)	48.5 (42.7, 54.4)

Table 1: **Hidden diagnosis rates across demographic subgroups.** We report the percentage of patients with a recorded diagnosis of AF across patients with a confirmed positive AF ECG (n=9,518), and patients with both a confirmed positive AF ECG and a visit to a primary care physician in the past year (n=2,459). AF diagnoses are present less often among racial minorities, Medicaid patients, and non-English speakers.

available physiological signal, addresses this limitation by enabling analysis of a substantially larger and more diverse cohort.

Deep learning model accurately predicts risk of AF across patient groups. To capture patient risk of AF, we train a deep convolutional neural network to distinguish positive examples (patients with a sinus rhythm ECG followed by an AF ECG within 90 days) from negative examples (patients with a sinus rhythm ECG and *no* AF ECG within 90 days). We train the model using standard techniques, as detailed in Methods.

The developed model predicts the probability of AF within 90 days as a function of morphological features extracted from a sinus rhythm ECG and demographics, and is able to effectively discriminate between patients who experience an AF ECG within 90 days and those who do not (achieving an AUC of 0.88). The model's performance is substantially higher than that of simpler models (e.g., a logistic regression) and simpler features (e.g., automatically extracted features of the ECG) (Figure 5). These results agree with prior literature showing that deep learning more effectively predicts AF compared to prior approaches [36, 37, 38, 39, 40], and the broader ability of deep neural networks to extract information from rich, physiologic data [41, 42]. Concretely, 9.6% of patients in the highest quintile of estimated risk experience AF within 90 days, while 0.2% of patients in the lowest quintile experience the same outcome. Risk estimates produced by the model also demonstrate that patient subgroups exhibit meaningfully different distributions in AF risk, suggesting the importance of controlling for risk when measuring diagnostic disparities (Figure 6). Controlling for risk ensures that measured disparities correspond to differences in diagnosis rates, rather than differences in underlying risk.

Importantly, the risk score performs similarly across salient demographic subgroups, achieving AUCs between 0.84 and 0.93 across patients of different races, insurance, and primary spoken languages (Figure 2). The risk score is also well-calibrated, as the model's predicted event rates closely align with empirical outcome rates within each group, despite substantial variation in group-specific event rates. We also find that the risk score is well-calibrated across the range of risk score values (Figure 7). Calibrated risk scores imply that conditional on risk score, we expect similar rates of AF ECGs across groups.

AF diagnosis rates are lower among underserved patients. We find substantial gaps in diagnosis rates between groups, conditional on risk (Figure 3). For example, the AF diagnosis rate among White patients in the highest quintile of risk (18.1%) is significantly higher than that among Black patients (14.8%, $p < 0.05$), Hispanic/Latino patients (12.3%, $p < 0.01$), and Asian patients (10.9% , $p < 0.01$). We also observe disparities in diagnosis by insurance and primary spoken language. Controlling for age, sex, site, health care utilization, and estimated risk, we estimate that absolute rates of diagnosis among Black patients are 1.8% (95% CI: 1.1%, 2.5%) lower than white patients, or 24% lower in relative terms. Absolute AF diagnosis rates are similarly lower among Hispanic/Latino and Asian patients relative to white patients (reductions of 1.5% (95% CI: 0.75%, 2.3%) and 2.7% (95% CI: 1.7%, 3.8%) respectively, Figure 4). We find lower rates of diagnosis conditional on risk among non-English speakers compared to English speakers, with an estimated 1.5% (0.78%, 2.3%) absolute reduction in diagnosis rate conditional on risk. Compared to commercially insured patients, Medicaid patients also had lower diagnosis rates, although the estimated disparity is not statistically significant. We estimate higher diagnosis rates for Medicare patients relative to commercially insured patients, potentially explained by widely-documented financial incentives to assign diagnoses more aggressively in Medicare Advantage [43, 44, 45, 46].

Diagnosis-based estimates of AF prevalence underestimate true prevalence among minorities. The regression results in Figure 4 suggest that patients of color, Medicaid-insured patients, and those primarily speaking a language other than English have disproportionately higher rates of hidden diagnosis. The regression model also allows us to ask: how many more diagnoses would we observe in a subgroup (e.g., Black patients), if they were diagnosed at the same rate as the reference group (e.g., white patients), conditional on risk? While we cannot know the true prevalence of AF among the reference group without ground truth labels, the regression model allows us to understand *changes* in estimated prevalence (e.g., that the estimated prevalence would increase or decrease if diagnosis rates were standardized to match the reference group). When diagnosis is not perfect in the reference group, as Table 1 reports, it implies that the true prevalence among the minority group is even higher. Using this approach, the regression model suggests that we would observe a 50% increase in AF diagnoses among minority patients (Table 2), had they been subject to the diagnosis rate of white patients, translating to 2,691 additional diagnoses between 2016 and 2019. Our estimates further suggest that approximately one-third of the gap in observed diagnosis rates between Black patients (3.5%) and white patients (7.9%) can be explained by hidden diagnosis.

Results are consistent across several robustness checks. Groups with lower diagnosis rates might actually be healthier in ways the risk score does not capture; i.e., the difference in diagnosis rates could be justified if, for example, Black patients are healthier than white patients for the same estimated risk. We check for this possibility in several ways. First, we measure differences in rates of AF complications (as measured by the presence of an ECG with a heart rate above 160, representing tachyarrhythmias, in Figures 3 and 4) conditional on estimated risk of AF. While we find substantial disparities in diagnosis rates conditional on risk, we find no differences in *complication* rates conditional on risks. If anything, these groups are sicker, and so differences in patient health do not explain differences in observed diagnosis rates. Results for additional AF complications are consistent and can be found in the Figure 8. Second, there are multiple plausible ways to define AF risk for a patient. We report robustness of our results to alternate risk score specifications (including different inclusion criteria and diagnosis definitions) and controls in Figure 10. Finally, we replicate our analysis among patients with no AF ECG within 90 days, demonstrating that our approach permits the estimation of diagnostic disparities even for patients who lack any downstream indicator of AF risk. Estimates of diagnosis rates are directionally consistent across all checks. Furthermore, a majority of differences in diagnosis rates remain significant across all robustness checks.

3 Discussion

This study demonstrates that AF diagnosis, despite its clinical importance and straightforward ECG presentation, is systematically "hidden" for traditionally underserved patient groups. We find clinically meaningful disparities in recorded diagnoses that are not attributable to differences in underlying risk by using machine learning to estimate and adjust for different distributions of disease risk between groups. While our study focuses on AF, it is plausible that hidden diagnosis occurs with other conditions as well, especially since AF is more straightforward to diagnose than many other conditions. One could apply the proposed framework to measure hidden diagnosis for other outcomes that are readily detectable from a downstream ECG (including, for example, other arrhythmias or structural heart conditions). Disparities in hidden diagnoses could directly affect patient care, ranging from missed opportunities for therapeutic intervention to misrepresentation in downstream systems that rely on recorded diagnoses. These biases, left unaddressed, could reinforce systematic inequalities at scale [47]. As a result, our findings suggest the need for systemic improvements in diagnostic and documentation practices and an urgent need for diagnosis-based research to develop methods to account for bias from hidden diagnosis.

Our findings have important implications for the use of health data. Diagnosis-based surrogates of atrial fibrillation will miss a large proportion of cases among underserved groups. Algorithms trained on such data risk encoding the very disparities we document. Furthermore, our framework provides actionable paths forward. For example, allocating diagnostic resources on the basis of estimated AF risk has been shown to double case detection compared to routine care in a clinical trial [48]; our results suggest it could also be an effective way to remedy diagnostic disparities. More broadly, diagnosis codes should be treated not as definitive labels but as imperfect, socially mediated signals that require auditing. Our approach illustrates one such diagnostic audit, offering health systems a way to identify and correct hidden biases.

Our results also contribute to the longstanding debate over why the observed prevalence of AF is consistently reported to be lower among Black patients than among white patients, despite comparable or higher risk factors in the former group[49, 50, 51, 52, 53, 54, 34, 35, 55, 56]. Prior work has variously attributed this gap to differences in genetics [53, 54], disease etiology [57], or documentation [34, 55, 56]. Our approach directly isolates differences in AF prevalence from differences in AF documentation, and suggests that both contribute to the lower observed rates of AF among Black patients. While we find that reporting differences are an important contributor in our sample, future work should strive to measure the contributions of each mechanism across diverse settings.

Several mechanisms could underpin rates of hidden diagnoses among underserved patients. One explanation could be fragmentation of care: patients can receive ECG evaluations disconnected from their primary health care providers, such that no provider assumes responsibility for recording diagnoses into structured EHR fields. A second issue can magnify this problem, which is that underserved patient groups may have lower health literacy [58] and therefore be less likely to advocate for documenting new conditions or recognize their absence in the chart. A third explanation is that underserved patient groups are less likely to present to clinicians for any reason [59, 20], potentially delaying or missing the opportunity to accurately code new conditions. Another possibility is implicit biases and differences in provider-patient interactions across demographic lines that could contribute to under-documentation. Although we include controls for measures of health care utilization, differences in health access beyond those we control for remain potential contributors. This list of explanations is not mutually exclusive and could contribute in combination to the results we find.

Our findings are subject to limitations. First, the study is confined to a single large academic health system. Although diverse, its practices and patient populations may not represent the broader healthcare landscape. Future work should validate these findings across multiple, varied healthcare systems to confirm generalizability. Second, our analysis excludes patients with little to no access to the healthcare system because it relies on access to a recorded ECG. However, this exclusion likely causes us to underestimate the full extent of disparities, as limited access to care may further compound the dispar-

		Observed Diagnosis Rate		Estimated Diagnosis Rate		Change in Diagnosis Rate	
		%	N	%	N	%	N
Race	Black or African American	3.4	1,202	5.0	1,782	48.3	580
	Hispanic or Latino	2.9	900	4.1	1,283	42.6	383
	Asian	3.3	589	5.5	921	56.4	332
	White	-	-	-	-	-	-
Insurance	Medicare	12.4	16,102	9.6	12,430	-22.8	-3,672
	Medicaid	2.4	986	2.6	1,053	6.8	67
	Commercial	-	-	-	-	-	-
Primary Language	Non-English	5.7	1,961	6.0	2,110	7.6	149
	English	-	-	-	-	-	-

Table 2: **Estimated prevalence in subgroups, using diagnosis rate of reference group.** The risk-adjusted regression framework allows us to estimate the true outcome prevalence for each demographic group: capturing what the diagnosis rate would have been in each patient subpopulation, had they experienced the diagnosis rate of the reference group (i.e. white patients, the commercially insured, and English speakers for race, insurance, and primary language respectively). The estimated prevalence of AF exceeds the observed prevalence of AF among racial minorities, those insured by Medicare, and non-English speakers. If Medicare patients experienced the diagnosis rate of the reference group (the commercially insured), prevalence of AF would decrease; we hypothesize that this effect could be due to increased diagnostic coding among Medicare beneficiaries relative to the commercially insured.

ities we observe. Further, our methods do not distinguish between different subtypes of AF, including paroxysmal or permanent AF and atrial flutter, or differing causes of known AF complications (rapid heart rate, stroke) that are not exclusively caused by AF, limiting granularity regarding clinical implications of missed diagnoses. Finally, our estimate of risk relies on the use of AF ECGs. Although we test the robustness of our findings to different risk score specifications in the supplement, advancements in our ability to estimate disease risk from physiological signals would increase the value of the proposed framework.

From a methodological standpoint, our work illustrates the utility of machine learning as not only a predictive tool but also an instrument for measuring biases inherent in health data [60, 61, 20, 62, 63, 64, 65, 66, 67]. Predictive algorithms can be used by health systems to assess hidden diagnoses in their own systems , and the underlying approach is a broadly applicable method, readily adaptable to other clinical conditions and health systems. The proposed approach represents a step towards health systems that can accurately locate and mitigate health inequities.

Author contributions MB, EP, and DS conceived the project. DS led the methodological development, experimental design, and data analysis. BH and AW assisted with data curation. BH, AW, SD, and MB contributed to the clinical interpretation of the results. All authors participated in writing and revising the manuscript and approved the final version. This study was reviewed and approved by the Institutional Review Board at Mass General Brigham under protocol number 2022P002371. All methods were carried out in accordance with relevant guidelines and regulations.

Materials & Correspondence All requests for materials and correspondence can be directed to divyas@cornell.edu. The data used in this study are not publicly available due to privacy constraints. To facilitate reproducibility, we will provide a synthetic dataset that mirrors the key statistical properties

of the original data and supports replication of the primary analyses upon publication. All code associated with the main and supplementary analyses is available at `https://github.com/divyashan/hidden_diagnosis/`.

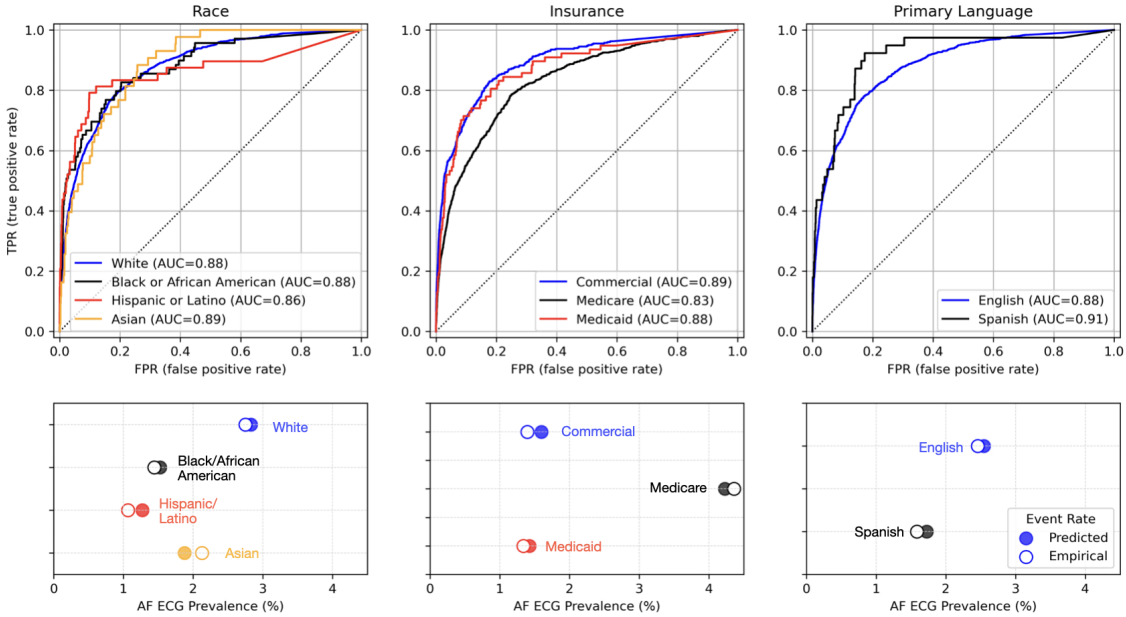

Figure 2: **Model validation.** The model of AF risk achieves high discriminative performance (an overall AUC of 0.88, with group-specific AUCs ranging between 0.84 - 0.93). Panels in the top row report model AUC across subgroups considered in this work. We measure the calibration of the model by comparing estimated event rates within a group (by averaging the predicted risk over patients in that group) to empirical event rates within the group. A well-calibrated model's estimated prevalence within demographic groups closely tracks empirical event rates, as the deep learning model does across demographic groups of interest (race, insurance, and primary spoken language, bottom row).

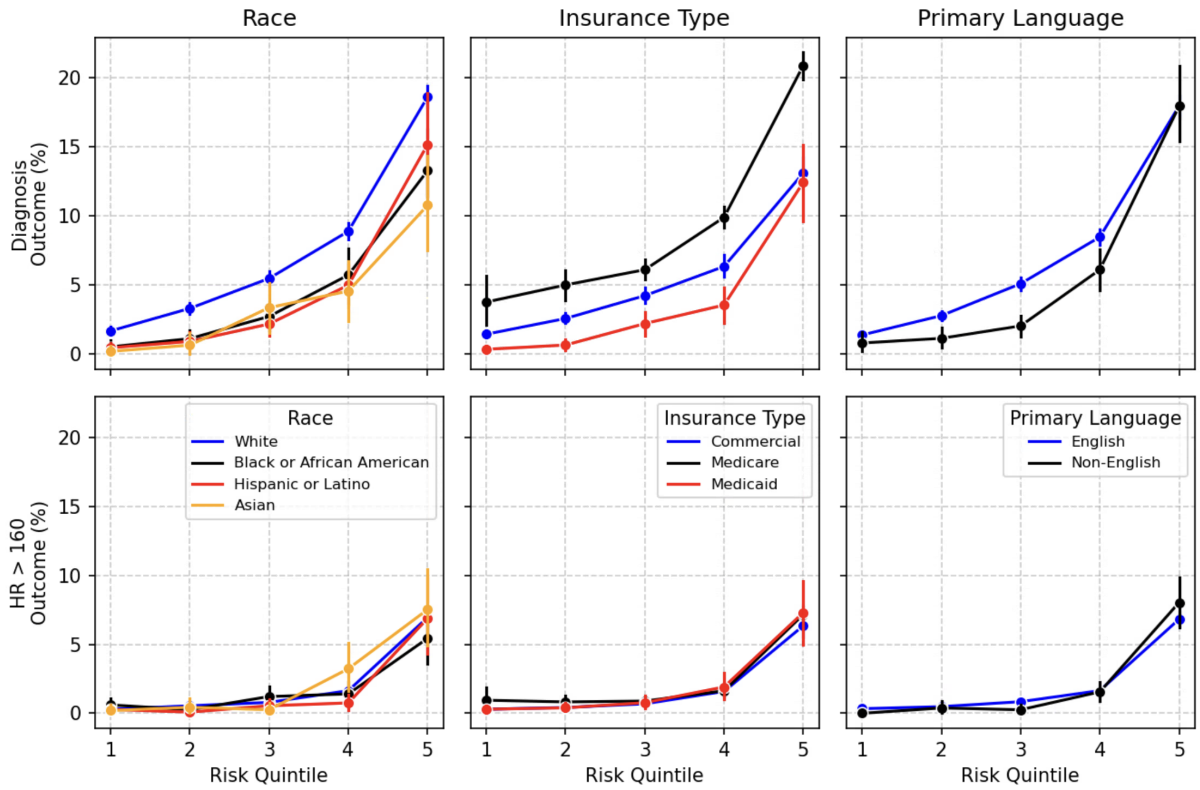

Figure 3: **Rates of AF diagnosis differ across demographic groups within a risk quintile, while rates of AF complications do not.** We plot rates of AF diagnosis (top row) by race, insurance, and primary language (left to right), comparing rates across risk quintiles (x-axis). Among race groups, white patients experience the highest rates of diagnosis conditional on risk. Similar trends hold in the context of insurance and primary language: patients with commercial insurance experience higher diagnosis rates than those on Medicaid (and lower than those on Medicare), and English-speakers experience higher diagnosis rates than those who do not speak English as a primary language. As a robustness check, we assess differences in rates of AF complications (e.g., an ECG with heart rate > 160 within a year, a clear marker of tachyarrhythmia designed to minimize false positives; bottom row) conditional on risk and find them to be negligible. Results for two additional AF complications are consistent can be found in the supplement. Estimated differences in diagnosis rates controlling for additional covariates (i.e. age, sex, site, and health access) can be found in Fig. 4.

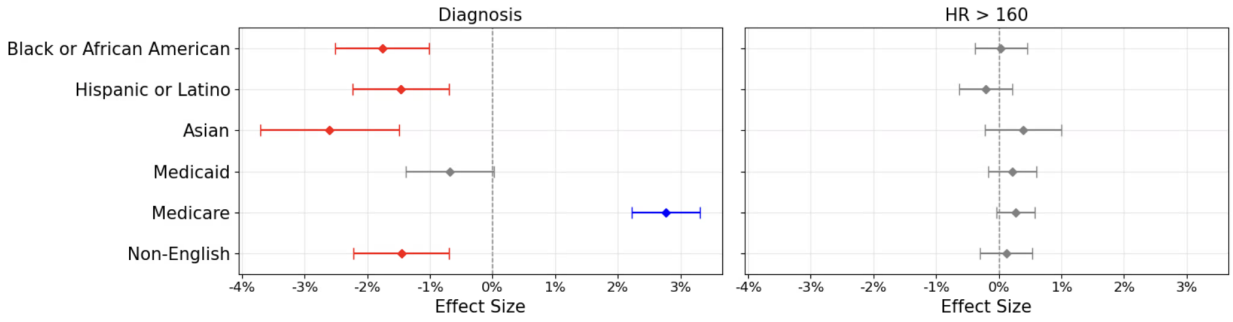

Figure 4: **Estimated differences in diagnosis rates across demographic groups**. We fit a linear probability model to recorded diagnosis, controlling for age, sex, and three binary indicators of healthcare access in the prior year. Diagnosis rates are significantly lower among Black, Hispanic/Latino, and Asian patients, along with non-English speakers. We also find higher rates of diagnosis among Medicare patients. The reference groups for race, insurance, and primary language are white patients, commercially insured patients, and English-speaking patients, respectively (where coefficients correspond to the absolute shift in diagnosis rate, relative to the reference group, left). We include coefficients associated with each demographic group when regressing an ECG with a heart rate greater than 160 within one year (right) to verify that the groups do not significantly vary in terms of more objective health measures, conditional on risk.

4 Data

While the data used in this study cannot be made publicly available due to privacy constraints, we will provide a synthetic dataset that mirrors key statistical properties of the original data to facilitate reproducibility. All code associated with the main and supplementary analyses is currently available at `https://github.com/divyashan/hidden_diagnosis/`.

4.1 Inputs

We use ECGs collected at 11 sites in a large academic health system between 2016 and 2019. We define an AF ECG as any 12-lead ECG with a physician reading of atrial fibrillation (based on a previously published list of AF findings [68]). We reproduce the complete list of findings below, for ease of reference. AF ECGs serve as a source of ground truth, and we will refer to patients with an AF ECG as patients with observed AF. For patients with no observed AF, their first sinus rhythm ECG serves as a negative example (for detail on sinus rhythm identification procedure, see Sec. 4.1.2). For patients with observed AF, we first identify the earliest ECG in which AF occurs, and retain patients who have a recorded sinus rhythm ECG in the 90 days prior to their AF ECG; these patients serve as positive examples. After filtering both the positive and negative samples for patients with complete demographic data and no prior AF ICD-10 diagnosis (i.e. an AF diagnosis that occurs prior to a patient's sinus rhythm ECG), the sample contains 406035 patients.

The resulting dataset thus captures a set of ECGs where (1) no AF occurs within 90 days (negative examples, n=396,517) and (2) AF occurs within 90 days (positive examples, n=9,518). We focus on the 90-day time horizon in accordance with prior work [39] and provide an illustration of our dataset creation procedure in Figure 1A. We follow standard ECG preprocessing procedures by removing baseline wander, truncating outliers by clipping feature values to the 0.1th–99.9th percentile range, and applying standard normalization of ECG values using mean and standard deviations computed on the training split. We reserve 60% of the ECG data for model training (n=243,824), 25% of the data for model calibration (n=101,381), and 15% of the data for our study split (n=60,830), split randomly at the patient level (i.e. no patient appears in any two splits).

4.1.1 Procedure to identify AF ECGs

We deem an ECG to exhibit atrial fibrillation if any of the following terms appear in the ECG's findings: "atrial fibrillation with rapid ventricular response", "atrial fibrillation with moderate ventricular response", "fibrillation/flutter", "atrial fibrillation with controlled ventricular response", "afib", "atrial fib", "afibrillation", "atrial fibrillation", or "atrialfibrillation" . Our procedure borrows directly from the Machine Learning for Cardiology and Critical Care pipeline developd by the Aguirre lab.

4.1.2 Procedure to identify sinus rhythm ECGs

We deem an ECG in sinus rhythm if any of the following keywords appear in the ECG's findings: "type i sinoatrial block", "conducted sinus impulses", "marked sinus arrhythmia", "normal when compared with ecg of", "rhythm remains normal sinus", "normal sinus rhythm", "frequent native sinus beats", "normal ecg", "atrialbigeminy", "sinoatrial block, type ii", "type ii sinoatrial block", "type ii sa block", "type i sa block", "sa block", "atrial trigeminy", "rhythm has reverted to normal", "rhythm is now clearly sinus", "sinus exit block", "tracing is within normal limits", "1st degree sa block", "sinus arrhythmia", "2nd degree sa block", "sinus tachycardia", "sinus rhythm at a rate", "sinus rhythm", "tracing within normal limits", "sinus mechanism has replaced", "atrial bigeminal rhythm", "sa exit block", "sinoatrial block", "rhythm is normal sinus", "with occasional native sinus beats", "sa block, type i", "sinus slowing", "atrial bigeminal rhythm", "atrial bigeminy and ventricular bigeminy", "sinus bradycardia". We source these keywords from the Machine Learning for Cardiology and Critical Care pipeline.

4.2 Demographics

We additionally extract information on each patient's age, sex, race, insurance (commercial, Medicaid, or Medicare), primary language (English or non-English), and measures of outpatient health access from the electronic health record. We do not include patients who could not be linked to demographic data (1.1% of sample) and patients for whom race data is unavailable (4.1% of sample). Age is calculated as the age at the time of sinus rhythm ECG collection using each patient's date of birth. Race and ethnicity data is self-reported and collected during patient registration. In cases where multiple race indicators are present for a single patient, we choose the first reported race for each patient. We define a patient's insurance as their primary effective insurance at the time their sinus rhythm ECG was recorded, and we treat patients enrolled in both Medicaid and Medicare as Medicare patients to simplify analysis. We also capture the main hospital affiliation in the health system according to where the ECG was captured (11 different sites). Finally, we extract three binary measures of health access for each patient corresponding to completion of a visit to a primary care physician, cardiologist, or any other specialist within the previous year. We define a qualifying primary care visit as an appointment or office visit within one of the following specialty departments: Internal Medicine, Family Medicine, Gerontology, and Primary Care. We define a qualifying cardiology visit as an office visit or appointment in the cardiology department. Finally, office visits or appointments with other departments qualify as visits to other specialists. These binary indicators serve as markers of both overall health and outpatient engagement in the health system.

4.3 Outcomes

We define a hidden diagnosis as the absence of a structured AF diagnosis –specifically, the absence of any ICD code from a broad set indicating atrial fibrillation or atrial flutter. We focus on structured diagnoses for two reasons. First, although AF may appear in clinical notes or ECG reports, these sources are rarely used in routine practice to establish the diagnosis: doing so would require manually reading each ECG report or searching free text for mentions of AF, actions that typically occur only when the clinician already suspects the condition. Second, ICD-based diagnoses are the standard surrogate for AF in the clinical and epidemiological literature [69, 70, 71] and form the primary signal used by algorithms to identify positive cases. For these reasons, we treat an AF diagnosis as "hidden" (both clinically and algorithmically) when it is not observable through the structured diagnoses.

To operationalize this definition, we extract AF diagnoses for all patients between January 1, 2016 and December 31, 2023. We define an AF diagnosis as the presence of any ICD code related to atrial fibrillation or atrial flutter, following prior work [72, 56]. We intentionally use a generous set of ICD codes to capture all plausible codes indicating AF. Our main analyses adopt an inclusive diagnosis window—any time prior to the end of the study period—ensuring that each patient has at least three years in which a diagnosis could be recorded (see Figure 10 for alternative definitions).

We also extract data on two AF complications: the occurrence of an ECG with heart rate above 160 (RVR), and stroke within one year (Stroke). We threshold heart rate at 160 to produce a clear marker of tachyarrhythmia and minimize the presence of false positives. We limit both outcomes to occurrences within a year of the sinus rhythm ECG date, and define stroke as the occurrence of a stroke-related diagnosis code in either the problem list or encounters data.

5 Methods

To create the AF risk score, we trained a deep convolutional neural network to predict a patient's probability of AF within 90 days (defined by physician ECG reading) based on a sinus rhythm ECG and demographics. The training sample included 243,824 patients (73% white, 8.8% Black, 8.3% Hispanic/Latino, 53% female, 89% English-speaking), with ECGs from January 1, 2016 to December 31, 2019 , where 1-2% of patients across subgroups had an AF ECG within 90 days. On the study split (n=60,830), the model

achieved an AUC of 0.87 and similar calibration across race groups. We then used a linear probability model to measure differences in diagnosis rates controlling for AF risk and observable covariates in the study split.

5.1 Training

The target label the network is trained to predict is whether a patient will experience atrial fibrillation (AF) within 90 days. We use a deep learning architecture consisting of 2.8 million parameters that was designed to extract information from raw ECGs [73] and initialize the network with weights trained to predict major adverse cardiovascular events. The network is trained for 10 epochs (with a learning rate of 1e-3, a batch size of 32, optimized using Adam). We select the model that achieves the highest AUC on the calibration set. While prior models to predict AF from an ECG exist [74, 75, 76, 39], we elect to train our own to ensure that patients in the study split do not appear in the data used to train the model.

5.2 Calibration

Once the model is trained, we adjust the risk score to ensure calibration across salient demographic subgroups using the calibration split (n=101,381). We do so by fitting a linear probability model to predict the probability of an AF ECG within 90 days from the risk score, age, sex, race, insurance, primary spoken language, and three binary measures of health access (presence of a visit to a primary care physician, cardiologist, or other specialist within the past year). We assess the calibration of predictions across groups by using the Spiegelhalter test [77], which compares the predicted and observed number of events under a binomial model and tests for significant differences between the two. We apply the Bonferroni correction [78] for multiple hypothesis testing because we test for miscalibration across 9 groups. Across all groups, we find that differences between predicted and observed event rates are not significant, representing no evidence of miscalibration.

5.3 Validation of risk score

We first assess the predictive performance of the developed model by measuring discriminative performance (AUC) and calibration across demographic groups (Figure 2), two standard metrics in prior work [39]. We next verify the utility of deep learning through comparisons to a logistic regression fit to patient demographics, a logistic regression fit to features extracted from the ECG, and a logistic regression fit to both demographics and preprocessed ECG features (Figure 5). While we cannot directly compare to CHARGE-AF model [79] due to the data available at the time of the ECG, a recent meta-analysis observed that such risk scores achieve AUCs of approximately 0.71 (with a 95% CI between 0.66 and 0.76) [80], substantially lower than that of the deep learning model.

5.4 Measuring diagnosis rates

To measure hidden diagnosis, we first apply the calibrated model to the study split of 60,830 patients to obtain patient-specific estimates of AF risk. Our goal is to measure differences in diagnosis rates between groups, conditional on estimated risk. We employ a linear probability model to estimate the probability of diagnosis conditional on a set of covariates, as is standard in risk-adjusted regression [21, 81]. Note that the linear probability model (LPM) is distinct from a logistic regression in that estimated coefficients correspond directly to marginal effects on the probability of the outcome. Specifically, we fit a linear probability model to predict the probability of an EHR diagnosis from the patient's risk score, race, insurance, and primary language, and control for the following factors: age, sex, site, and binary measures of health access. We control for site to adjust for variation in patient demographics and coding practices across sites, a pattern well-established by the health disparities literature [82, 83, 84, 85]. Each binary

$$diagnosis \sim risk \ + \ age \ + \ sex \ + \ site + \ visit_{PCP} \ + visit_{Spec} \ + visit_{Oth} \ + \ race \tag{1}$$

$$diagnosis \sim risk \ + \ age \ + \ sex \ + \ site + visit_{PCP} \ + visit_{Spec} \ + visit_{Oth} \ + \ ins \tag{2}$$

$$diagnosis \sim risk \ + \ age \ + \ sex \ + \ site + \ visit_{PCP} \ + visit_{Spec} \ + visit_{Oth} \ + \ language \tag{3}$$

measure of health access captures the presence of a visit to a PCP, specialist, or other doctor within the past year. We use these LPMs to arrive at demographic-specific estimates of diagnosis rates relative to the reference group (where the reference groups for race, insurance, and primary language are white patients, the commercially insured, and English speakers).

5.5 Estimating hidden diagnoses

One can use the linear probability model to infer the number of diagnoses that would have occurred, had a group been subject to the same diagnosis rate as the reference group. For example, to estimate the number of hidden diagnoses for each racial minority, we can apply the linear probability model to each patient in a particular racial group, zeroing out the race variable to apply the diagnosis rate of the reference group. We perform this procedure for each demographic group separately to arrive at the number of hidden diagnoses within each group.

5.6 Sensitivity analyses

We test the robustness of our results across seven sensitivity analyses, detailed below. The sensitivity analyses probe the robustness of our findings in three categories: risk score quality, dataset construction, and regression specification.

5.6.1 Risk score quality

The learned coefficients estimate the difference in diagnosis rates, relative to the reference group. To verify that these effects are not due to group-specific differences in underlying risk, we additionally fit a linear probability model to three outcomes: an AF positive ECG within 90 days, an ECG with a heart rate 160 (an AF complication), and stroke within 1 year of the sinus rhythm ECG (another AF complication). We do then test that outcome rates do not significantly differ between groups, conditional on risk.

5.6.2 Dataset construction

The construction of our dataset implicitly assumes that patients in different demographic groups are equally likely to receive an ECG within 90 days, and thus have an ECG in AF observed within the qualifying window. We find that our results hold under three sensitivity analyses which each aim to control for variation in health access. First, we train the risk score on the single largest race group (white patients), to control for known variation in health access across racial groups [86, 18]. Second, we train the risk score by restricting negative examples to those where we observe a second sinus rhythm ECG within the prediction window. Both the direction and magnitude of estimated coefficients are consistent across both analyses. Further, violations of the assumption would lead us to *underestimate* the magnitude of disparities in diagnosis. For patients who are less likely to receive a second ECG, predicted risk would underestimate true risk. Thus, conditional on risk, underserved patients are expected to be less healthy compared to the reference group, which would *reduce* the observed gaps in diagnosis rates conditional on risk. Finally, we replicate our analysis on patients with no downstream indication of risk, i.e. no AF ECG within 90 days. Here too, we see consistent estimates of diagnostic disparities.

5.6.3 Regression specification

Finally, we report coefficient estimates when controlling for all demographic variables (which we term the joint regression), along with coefficient estimates under different time horizons for diagnosis (6 months, 1 year, 2 years, and 3 years).

References

[1] Crystian B Oliveira, Chris G Maher, Rafael Z Pinto, Adrian C Traeger, Chung-Wei Christine Lin, Jean-François Chenot, Maurits Van Tulder, and Bart W Koes. Clinical practice guidelines for the management of non-specific low back pain in primary care: an updated overview. *European Spine Journal*, 27:2791–2803, 2018.

[2] Stuart H Ralston, Luis Corral-Gudino, Cyrus Cooper, Roger M Francis, William D Fraser, Luigi Gennari, Núria Guanabens, M Kassim Javaid, Robert Layfield, Terence W O'Neill, et al. Diagnosis and management of paget's disease of bone in adults: a clinical guideline. *Journal of bone and mineral research*, 34(4):579–604, 2019.

[3] Peter S Hussey, Han De Vries, John Romley, Margaret C Wang, Susan S Chen, Paul G Shekelle, and Elizabeth A McGlynn. A systematic review of health care efficiency measures. *Health services research*, 44(3):784–805, 2009.

[4] Katherine J Hoggatt, Alex HS Harris, Corey J Hayes, Donna Washington, and Emily C Williams. Improving diagnosis-based quality measures: an application of machine learning to the prediction of substance use disorder among outpatients. *BMJ Open Quality*, 14(1), 2025.

[5] Valerie JM Watzlaf, Jennifer Hornung Garvin, Sohrab Moeini, and Patricia Anania-Firouzan. The effectiveness of icd-10-cm in capturing public health diseases. *Perspectives in Health Information Management/AHIMA, American Health Information Management Association*, 4:6, 2007.

[6] Jeremy A Rassen, Dorothee B Bartels, Sebastian Schneeweiss, Amanda R Patrick, and William Murk. Measuring prevalence and incidence of chronic conditions in claims and electronic health record databases. *Clinical epidemiology*, pages 1–15, 2018.

[7] Pandora L Wander, Aaron Baraff, Alexandra Fox, Kelly Cho, Monika Maripuri, Jacqueline P Honerlaw, Yuk-Lam Ho, Andrew T Dey, Ann M O'Hare, Amy SB Bohnert, et al. Rates of icd-10 code u09. 9 documentation and clinical characteristics of va patients with post–covid-19 condition. *JAMA Network Open*, 6(12):e2346783–e2346783, 2023.

[8] Marliese Alexander, Benjamin Solomon, David L Ball, Mimi Sheerin, Irene Dankwa-Mullan, Anita M Preininger, Gretchen Purcell Jackson, and Dishan M Herath. Evaluation of an artificial intelligence clinical trial matching system in australian lung cancer patients. *JAMIA open*, 3(2):209–215, 2020.

[9] Neha Jain, Kathleen F Mittendorf, Marilyn Holt, Michele Lenoue-Newton, Ian Maurer, Clinton Miller, Matthew Stachowiak, Michelle Botyrius, James Cole, Christine Micheel, et al. The my cancer genome clinical trial data model and trial curation workflow. *Journal of the American Medical Informatics Association*, 27(7):1057–1066, 2020.

[10] Gopi Battineni, Getu Gamo Sagaro, Nalini Chinatalapudi, and Francesco Amenta. Applications of machine learning predictive models in the chronic disease diagnosis. *Journal of personalized medicine*, 10(2):21, 2020.

[11] Yazeed Zoabi, Shira Deri-Rozov, and Noam Shomron. Machine learning-based prediction of covid-19 diagnosis based on symptoms. *npj digital medicine*, 4(1):1–5, 2021.

[12] Maria Ferris, Randal K Detwiler, Abhijit V Kshirsagar, Margareth Pierre-Louis, Lawrence Mandhelker, and David A Shoham. High prevalence of unlabeled chronic kidney disease among inpatients at a tertiary-care hospital. *The American Journal of the Medical Sciences*, 337(2):93–97, 2009.

[13] Stacey E Jolly, Sankar D Navaneethan, Jesse D Schold, Susana Arrigain, John W Sharp, Anil K Jain, Martin J Schreiber, James F Simon, and Joseph V Nally. Chronic kidney disease in an electronic health record problem list: quality of care, esrd, and mortality. *American journal of nephrology*, 39(4):288–296, 2014.

[14] Clarissa J Diamantidis, Sarah L Hale, Virginia Wang, Valerie A Smith, Sarah Hudson Scholle, and Matthew L Maciejewski. Lab-based and diagnosis-based chronic kidney disease recognition and staging concordance. *BMC nephrology*, 20:1–10, 2019.

[15] Jenna M Norton, Lindsay Grunwald, Amanda Banaag, Cara Olsen, Andrew S Narva, Eric Marks, and Tracey P Koehlmoos. Ckd prevalence in the military health system: coded versus uncoded ckd. *Kidney Medicine*, 3(4):586–595, 2021.

[16] T. Waidmann and S. Rajan. Race and ethnic disparities in health care access and utilization: an examination of state variation. *Medical care research and review : MCRR*, 57 Suppl 1:55–84, 2000.

[17] Kan Z Gianattasio, Christina Prather, M Maria Glymour, Adam Ciarleglio, and Melinda C Power. Racial disparities and temporal trends in dementia misdiagnosis risk in the united states. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 5:891–898, 2019.

[18] C. Caraballo, D. Massey, S. Mahajan, Yuan Lu, A. Annapureddy, B. Roy, C. Riley, K. Murugiah, J. Valero-Elizondo, O. Onuma, M. Nunez-Smith, H. Forman, K. Nasir, J. Herrin, and H. Krumholz. Racial and ethnic disparities in access to health care among adults in the united states: A 20-year national health interview survey analysis, 1999–2018. In *unknown*, 2020.

[19] E. Dotan, Shannon M Lynch, Joanne C Ryan, and Edith P Mitchell. Disparities in care of older adults of color with cancer: A narrative review. In *unknown*, 2024.

[20] Erica Chiang, Divya Shanmugam, Ashley N Beecy, Gabriel Sayer, Nir Uriel, Deborah Estrin, Nikhil Garg, and Emma Pierson. Learning disease progression models that capture health disparities. *arXiv preprint arXiv:2412.16406*, 2024.

[21] Jongbin Jung, Sam Corbett-Davies, Ravi Shroff, and Sharad Goel. Omitted and included variable bias in tests for disparate impact. *arXiv*, abs/1809.05651, 2018.

[22] Jean Jacques Noubiap, Janet J Tang, Justin T Teraoka, Thomas A Dewland, and Gregory M Marcus. Minimum national prevalence of diagnosed atrial fibrillation inferred from california acute care facilities. *Journal of the American College of Cardiology*, 84(16):1501–1508, 2024.

[23] P. Kirchhof, G. Breithardt, J. Bax, G. Benninger, C. Blomstrom-Lundqvist, G. Boriani, A. Brandes, Helen Brown, M. Brueckmann, H. Calkins, M. Calvert, V. Christoffels, H. Crijns, D. Dobrev, P. Ellinor, L. Fabritz, T. Fetsch, S. B. Freedman, A. Gerth, A. Goette, E. Guasch, Guido Hack, L. Haegeli, S. Hatem, K. Haeusler, H. Heidbüchel, J. Heinrich-Nols, F. Hidden-Lucet, G. Hindricks, S. Juul-Möller, S. Kääb, L. Kappenberger, S. Kespohl, D. Kotecha, Deirdre A Lane, A. Leute, T. Lewalter, Ralf Meyer, L. Mont, Felix Münzel, M. Nabauer, J. C. Nielsen, M. Oeff, J. Oldgren, A. Oto, J. Piccini, A. Pilmeyer, T. Potpara, U. Ravens, H. Reinecke, T. Rostock, Joerg Rustige, I. Savelieva, R. Schnabel, U. Schotten, L. Schwichtenberg, Moritz F. Sinner, G. Steinbeck, M. Stoll, L. Tavazzi, S. Themistoclakis, H. Tse, I. V. Van Gelder, P. Vardas, T. Varpula, Alphons Vincent, D. Werring, S. Willems, A. Ziegler, G. Lip, and A. Camm. A roadmap to improve the quality of atrial fibrillation management: proceedings from the fifth atrial fibrillation network/european heart rhythm association consensus conference. *Europace : European pacing, arrhythmias, and cardiac electrophysiology : journal of the working groups on cardiac pacing, arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology*, 18 1:37–50, 2016.

[24] N. Jones, C. Taylor, Hobbs Fdr., L. Bowman, and B. Casadei. Screening for atrial fibrillation: a call for evidence. *European Heart Journal*, 41:1075 – 1085, 2019.

[25] Josep L Clua-Espuny, Iñigo Lechuga-Duran, Ramón Bosch-Princep, Albert Roso-Llorach, Anna Panisello-Tafalla, Jorgina Lucas-Noll, Carles López-Pablo, Lluïsa Queralt-Tomas, Emmanuel Giménez-Garcia, Núria González-Rojas, et al. Prevalence of undiagnosed atrial fibrillation and of that not being treated with anticoagulant drugs: the afabe study. *Revista Española de Cardiología (English Edition)*, 66(7):545–552, 2013.

[26] Stefano Omboni and Willem J Verberk. Opportunistic screening of atrial fibrillation by automatic blood pressure measurement in the community. *BMJ open*, 6(4):e010745, 2016.

[27] Steven R Steinhubl, Jill Waalen, Alison M Edwards, Lauren M Ariniello, Rajesh R Mehta, Gail S Ebner, Chureen Carter, Katie Baca-Motes, Elise Felicione, Troy Sarich, et al. Effect of a home-based wearable continuous ecg monitoring patch on detection of undiagnosed atrial fibrillation: the mstops randomized clinical trial. *Jama*, 320(2):146–155, 2018.

[28] Giorgio Quer, Ben Freedman, and Steven R Steinhubl. Screening for atrial fibrillation: predicted sensitivity of short, intermittent electrocardiogram recordings in an asymptomatic at-risk population. *EP Europace*, 22(12):1781–1787, 2020.

[29] Nikolas Nozica, Anna Lam, Eleni Goulouti, Elena Georgieva Elchinova, Alessandro Spirito, Mattia Branca, Helge Simon Servatius, Fabian Noti, Jens Seiler, Samuel Hannes Baldinger, et al. The silent atrial fibrillation (star-fib) study programme-design and rationale. *Swiss medical weekly*, 151:w20421, 2021.

[30] Eiichi Watanabe, Naohiko Takahashi, Ronald Aronson, Ako Ohsawa, Yuriko Ishibashi, Yuji Murakawa, SCAN-AF Investigators, et al. Systematic screening for atrial fibrillation in patients at moderate-to-high risk of stroke—potential to increase the atrial fibrillation detection rate (scan-af)—. *Circulation Journal*, 86(8):1245–1251, 2022.

[31] Mintu P Turakhia, Jason Shafrin, Katalin Bognar, Jeffrey Trocio, Younos Abdulsattar, Daniel Wiederkehr, and Dana P Goldman. Estimated prevalence of undiagnosed atrial fibrillation in the united states. *PloS one*, 13(4):e0195088, 2018.

[32] Vegard Malmo, Arnulf Langhammer, Kaare H Bønaa, Jan P Loennechen, and Hanne Ellekjaer. Validation of self-reported and hospital-diagnosed atrial fibrillation: the hunt study. *Clinical Epidemiology*, pages 185–193, 2016.

[33] Mintu P Turakhia, Jennifer D Guo, Allison Keshishian, Rachel Delinger, Xiaoxi Sun, Mauricio Ferri, Cristina Russ, Matthew Cato, Huseyin Yuce, and Patrick Hlavacek. Contemporary prevalence estimates of undiagnosed and diagnosed atrial fibrillation in the united states. *Clinical Cardiology*, 46(5):484–493, 2023.

[34] Susan R Heckbert, Thomas R Austin, Paul N Jensen, Lin Y Chen, Wendy S Post, James S Floyd, Elsayed Z Soliman, Richard A Kronmal, and Bruce M Psaty. Differences by race/ethnicity in the prevalence of clinically detected and monitor-detected atrial fibrillation: Mesa. *Circulation: Arrhythmia and Electrophysiology*, 13(1):e007698, 2020.

[35] Utibe R Essien, Jelena Kornej, Amber E Johnson, Lucy B Schulson, Emelia J Benjamin, and Jared W Magnani. Social determinants of atrial fibrillation. *Nature Reviews Cardiology*, 18(11):763–773, 2021.

[36] B. Pourbabaee, M. J. Roshtkhari, and K. Khorasani. Deep convolutional neural networks and learning ecg features for screening paroxysmal atrial fibrillation patients. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48:2095–2104, 2018.

[37] Z. Attia, P. Noseworthy, F. Lopez-Jimenez, S. Asirvatham, A. Deshmukh, B. Gersh, R. Carter, Xiaoxi Yao, A. Rabinstein, Brad J Erickson, S. Kapa, and P. Friedman. An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet*, 394:861–867, 2019.

[38] S. Khurshid, S. Friedman, C. Reeder, P. Di Achille, N. Diamant, Pulkit Singh, Lia X. Harrington, Xin Wang, M. Al-Alusi, Gopal Sarma, A. Foulkes, P. Ellinor, C. Anderson, J. Ho, A. Philippakis, P. Batra, and S. Lubitz. Ecg-based deep learning and clinical risk factors to predict atrial fibrillation. *Circulation*, 145:122 – 133, 2021.

[39] Neal Yuan, Grant Duffy, Sanket S Dhruva, Adam Oesterle, Cara N Pellegrini, John Theurer, Marzieh Vali, Paul A Heidenreich, Salomeh Keyhani, and David Ouyang. Deep learning of electrocardiograms in sinus rhythm from us veterans to predict atrial fibrillation. *JAMA cardiology*, 8(12):1131–1139, 2023.

[40] M. Gadaleta, P. Harrington, Eric Barnhill, E. Hytopoulos, M. Turakhia, S. Steinhubl, and G. Quer. Prediction of atrial fibrillation from at-home single-lead ecg signals without arrhythmias. *NPJ Digital Medicine*, 6, 2023.

[41] Aly A. Valliani, D. Ranti, and E. Oermann. Deep learning and neurology: A systematic review. *Neurology and Therapy*, 8:351 – 365, 2019.

[42] Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, G. Tison, Codie Bourn, M. Turakhia, and A. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25:65 – 69, 2019.

[43] US Department of Health and Human Services. Medicare Advantage: Questionable Use of Health Risk Assessments Continues To Drive Up Payments to Plans by Billions, October 2024.

[44] Michael Geruso and Timothy Layton. Upcoding: evidence from medicare on squishy risk adjustment. *Journal of Political Economy*, 128(3):984–1026, 2020.

[45] Steven M Lieberman and Paul B Ginsburg. Improving medicare advantage by accounting for large differences in upcoding across plans. *Health Affairs Forefront*, 2025.

[46] Q: What can help us fund comprehensive care for our Medicare population? A: Use the EHR to improve HCC capture.

[47] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[48] Nathan R Hill, Lara Groves, Carissa Dickerson, Andreas Ochs, Dong Pang, Sarah Lawton, Michael Hurst, Kevin G Pollock, Daniel M Sugrue, Carmen Tsang, et al. Identification of undiagnosed atrial fibrillation using a machine learning risk-prediction algorithm and diagnostic testing (pulse-ai) in primary care: a multi-centre randomized controlled trial in england. *European Heart Journal-Digital Health*, 3(2):195–204, 2022.

[49] Elsayed Z Soliman, Alvaro Alonso, and David C Goff. Atrial fibrillation and ethnicity: the known, the unknown and the paradox. *Future cardiology*, 5(6):547–556, 2009.

[50] Alvaro Alonso, Sunil K Agarwal, Elsayed Z Soliman, Marietta Ambrose, Alanna M Chamberlain, Ronald J Prineas, and Aaron R Folsom. Incidence of atrial fibrillation in whites and african-americans: the atherosclerosis risk in communities (aric) study. *American heart journal*, 158(1):111–117, 2009.

[51] Elsayed Z Soliman and Ronald J Prineas. The paradox of atrial fibrillation in african americans. *Journal of Electrocardiology*, 47(6):804–808, 2014.

[52] Thomas D Stamos and Dawood Darbar. The "double" paradox of atrial fibrillation in black individuals. *JAMA cardiology*, 1(4):377–379, 2016.

[53] Gregory M Marcus, Alvaro Alonso, Carmen A Peralta, Guillaume Lettre, Eric Vittinghoff, Steven A Lubitz, Ervin R Fox, Yamini S Levitzky, Reena Mehra, Kathleen F Kerr, et al. European ancestry as a risk factor for atrial fibrillation in african americans. *Circulation*, 122(20):2009–2015, 2010.

[54] Jason D Roberts, Donglei Hu, Susan R Heckbert, Alvaro Alonso, Thomas A Dewland, Eric Vittinghoff, Yongmei Liu, Bruce M Psaty, Jeffrey E Olgin, Jared W Magnani, et al. Genetic investigation into the differential risk of atrial fibrillation among black and white individuals. *JAMA cardiology*, 1(4):442–450, 2016.

[55] Richard G Trohman, Henry D Huang, and Parikshit S Sharma. Atrial fibrillation: primary prevention, secondary prevention, and prevention of thromboembolic complications: part 1. *Frontiers in Cardiovascular Medicine*, 10:1060030, 2023.

[56] Lars Hulstaert, Amelia Boehme, Kaitlin Hood, Jennifer Hayden, Clark Jackson, Astra Toyip, Hans Verstraete, Yu Mao, and Khaled Sarsour. Assessing ascertainment bias in atrial fibrillation across us minority groups. *Plos one*, 19(4):e0301991, 2024.

[57] James F Meschia, Peter Merrill, Elsayed Z Soliman, Virginia J Howard, Kevin M Barrett, Neil A Zakai, Dawn Kleindorfer, Monika Safford, and George Howard. Racial disparities in awareness and treatment of atrial fibrillation: the reasons for geographic and racial differences in stroke (regards) study. *Stroke*, 41(4):581–587, 2010.

[58] Coraline Stormacq, Stephan Van den Broucke, and Jacqueline Wosinski. Does health literacy mediate the relationship between socioeconomic status and health disparities? integrative review. *Health promotion international*, 34(5):e1–e17, 2019.

[59] Michael L Barnett, Asaf Bitton, Jeff Souza, and Bruce E Landon. Trends in outpatient care for medicare beneficiaries and implications for primary care, 2000 to 2019. *Annals of internal medicine*, 174(12):1658–1665, 2021.

[60] Sendhil Mullainathan and Ziad Obermeyer. Diagnosing physician error: A machine learning approach to low-value health care. *The Quarterly Journal of Economics*, 137(2):679–727, 2022.

[61] Divya Shanmugam, Kaihua Hou, and Emma Pierson. Quantifying disparities in intimate partner violence: A machine learning method to correct for underreporting. *npj Women's Health*, 2(1):15, 2024.

[62] Emma Pierson. Assessing racial inequality in covid-19 testing with bayesian threshold tests. *NeurIPS Machine Learning for Health Workshop*, 2020.

[63] Keith Harrigian, Ayah Zirikly, Brant Chee, Alya Ahmad, Anne Links, Somnath Saha, Mary Catherine Beach, and Mark Dredze. Characterization of stigmatizing language in medical records. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–329, 2023.

[64] Emma Pierson, Divya Shanmugam, Rajiv Movva, Jon Kleinberg, Monica Agrawal, Mark Dredze, Kadija Ferryman, Judy Wawira Gichoya, Dan Jurafsky, Pang Wei Koh, Karen Levy, Sendhil Mullainathan, Ziad Obermeyer, Harini Suresh, and Keyon Vafa. Using large language models to promote health equity. *NEJM AI*, 2(2):AIp2400889, 2025.

[65] Kadija Ferryman, Maxine Mackintosh, and Marzyeh Ghassemi. Considering biased data as informative artifacts in ai-assisted health care. *New England Journal of Medicine*, 389(9):833–838, 2023.

[66] Emma Pierson, David M Cutler, Jure Leskovec, Sendhil Mullainathan, and Ziad Obermeyer. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, 27(1):136–140, 2021.

[67] Sidhika Balachandar, Nikhil Garg, and Emma Pierson. Domain constraints improve risk prediction when outcome data is missing. *ICLR*, 2024.

[68] Nathaniel Diamant, Erik Reinertsen, Steven Song, Aaron D Aguirre, Collin M Stultz, and Puneet Batra. Patient contrastive learning: A performant, expressive, and practical approach to electrocardiogram modeling. *PLoS computational biology*, 18(2):e1009862, 2022.

[69] Jared W Magnani, Faye L Norby, Sunil K Agarwal, Elsayed Z Soliman, Lin Y Chen, Laura R Loehr, and Alvaro Alonso. Racial differences in atrial fibrillation-related cardiovascular disease and mortality: the atherosclerosis risk in communities (aric) study. *JAMA cardiology*, 1(4):433–441, 2016.

[70] Premanand Tiwari, Kathryn L Colborn, Derek E Smith, Fuyong Xing, Debashis Ghosh, and Michael A Rosenberg. 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, 2020.

[71] Ryoko Suzuki, Jun Katada, Sreeram Ramagopalan, and Laura McDonald. Potential of machine learning methods to identify patients with nonvalvular atrial fibrillation. *Future Cardiology*, 16(1):43–52, 2020.

[72] Ren Jie Robert Yao, Jason G Andrade, Marc W Deyell, Heather Jackson, Finlay A McAlister, and Nathaniel M Hawkins. Sensitivity, specificity, positive and negative predictive values of identifying atrial fibrillation using administrative data: a systematic review and meta-analysis. *Clinical epidemiology*, pages 753–767, 2019.

[73] David Ouyang, John Theurer, Nathan R Stein, J Weston Hughes, Pierre Elias, Bryan He, Neal Yuan, Grant Duffy, Roopinder K Sandhu, Joseph Ebinger, et al. Electrocardiographic deep learning for predicting post-procedural mortality: a model development and validation study. *The Lancet Digital Health*, 6(1):e70–e78, 2024.

[74] Zachi I Attia, Peter A Noseworthy, Francisco Lopez-Jimenez, Samuel J Asirvatham, Abhishek J Deshmukh, Bernard J Gersh, Rickey E Carter, Xiaoxi Yao, Alejandro A Rabinstein, Brad J Erickson, 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. *The Lancet*, 394(10201):861–867, 2019.

[75] Shaan Khurshid, Samuel Friedman, Shinwan Kany, Jonathan Cunningham, Emily Lau, Daniel Pipilas, Mostafa Al-Alusi, Joel Ramo, James Pirruccello, Victor Nauffal, et al. Electrocardiogram-based artificial intelligence predicts incident heart failure. *Circulation*, 148(Suppl_1):A15965–A15965, 2023.

[76] Jagmeet P Singh, Julien Fontanarava, Grégoire de Massé, Tanner Carbonati, Jia Li, Christine Henry, and Laurent Fiorina. Short-term prediction of atrial fibrillation from ambulatory monitoring ecg using a deep neural network. *European Heart Journal-Digital Health*, 3(2):208–217, 2022.

[77] Yingxiang Huang, Wentao Li, Fima Macheret, Rodney A Gabriel, and Lucila Ohno-Machado. A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*, 27(4):621–633, 2020.

[78] Eric W Weisstein. Bonferroni correction. *https://mathworld. wolfram. com/*, 2004.

[79] Alvaro Alonso, Bouwe P Krijthe, Thor Aspelund, Katherine A Stepas, Michael J Pencina, Carlee B Moser, Moritz F Sinner, Nona Sotoodehnia, João D Fontes, A Cecile JW Janssens, et al. Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the charge-af consortium. *Journal of the American Heart Association*, 2(2):e000102, 2013.

[80] Jelle CL Himmelreich, Lieke Veelers, Wim AM Lucassen, Renate B Schnabel, Michiel Rienstra, Henk CPM van Weert, and Ralf E Harskamp. Prediction models for atrial fibrillation applicable in the community: a systematic review and meta-analysis. *EP Europace*, 22(5):684–694, 2020.

[81] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

[82] Sean P Clarke, Bertha L Davis, and Regina E Nailon. Racial segregation and differential outcomes in hospital care. *Western Journal of Nursing Research*, 29(6):739–757, 2007.

[83] Mary S Vaughan Sarrazin, Mary E Campbell, Kelly K Richardson, and Gary E Rosenthal. Racial segregation and disparities in health care delivery: conceptual model and empirical assessment. *Health services research*, 44(4):1424–1444, 2009.

[84] Michael B Rothberg, Penelope S Pekow, Aruna Priya, and Peter K Lindenauer. Variation in diagnostic coding of patients with pneumonia and its association with hospital risk-standardized mortality rates: a cross-sectional analysis. *Annals of internal medicine*, 160(6):380–388, 2014.

[85] Andrea M Austin, Donald Q Carmichael, Julie PW Bynum, and Jonathan S Skinner. Measuring racial segregation in health system networks using the dissimilarity index. *Social science & medicine*, 240:112570, 2019.

[86] Jie Chen, Arturo Vargas-Bustamante, Karoline Mortensen, and A. Ortega. Racial and ethnic disparities in health care access and utilization under the affordable care act. *Medical Care*, 54:140 – 146, 2016.

		Train	Calibration	Study
Overall	# of Patients	243,824	101,381	60,830
	Age (Mean (SD))	56.3 (18.1)	56.3 (18.2)	56.5 (18.1)
Sex	Male	114,761 (47.1%)	47,540 (46.9%)	28,445 (46.8%)
	Female	128,724 (52.8%)	53,799 (53.1%)	32,359 (53.2%)
Race	White	178,534 (73.2%)	74,488 (73.5%)	44,690 (73.5%)
	Black or African American	21,458 (8.8%)	8,986 (8.9%)	5,339 (8.8%)
	Hispanic or Latino	20,313 (8.3%)	8,334 (8.2%)	5,055 (8.3%)
	Asian	8,913 (3.7%)	3,763 (3.7%)	2,245 (3.7%)
Insurance	Commercial	119,571 (49.0%)	49,895 (49.2%)	29,991 (49.3%)
	Medicare	65,499 (26.9%)	27,158 (26.8%)	16,275 (26.8%)
	Medicaid	25,130 (10.3%)	10,426 (10.3%)	6,257 (10.3%)
Primary Language	English	217,733 (89.3%)	90,749 (89.5%)	54,363 (89.4%)
	Non-English	26,091 (10.7%)	10,632 (10.5%)	6,467 (10.6%)
Outcome	AF Diagnosis	16,229 (6.7%)	6,859 (6.8%)	4,018 (6.6%)
	AF ECG ($<$ 90 days)	5,748 (2.4%)	2,328 (2.3%)	1,442 (2.4%)
	Stroke	22,591 (9.3%)	9,363 (9.2%)	5,629 (9.3%)

Table 3: **Descriptive statistics for train, calibration, and study splits.** We report number of patients in different demographic categories, accompanied by the percentage of the split. The training and calibration split are used to train and calibrate the deep learning model, and we measure diagnostic disparities on the study split.

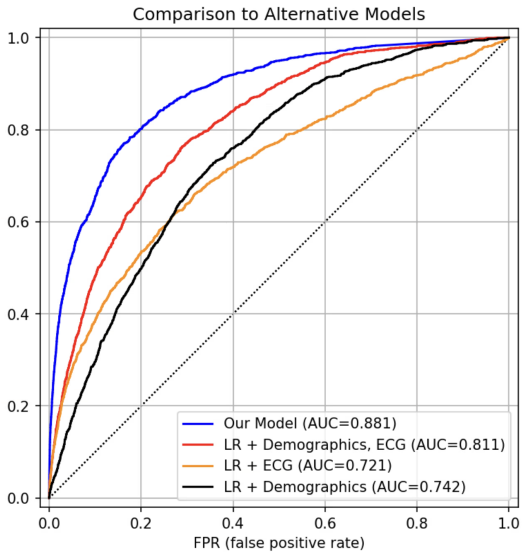

Figure 5: **Comparison to simpler models.** We compare performance of the deep learning model to three simpler models: a logistic regression fit to demographics alone (black), preprocessed ECG features alone (yellow), and a logistic regression fit to both sets of features (red). The deep learning model (blue) outperforms these simpler models by a large margin, achieving an AUC of 0.88.

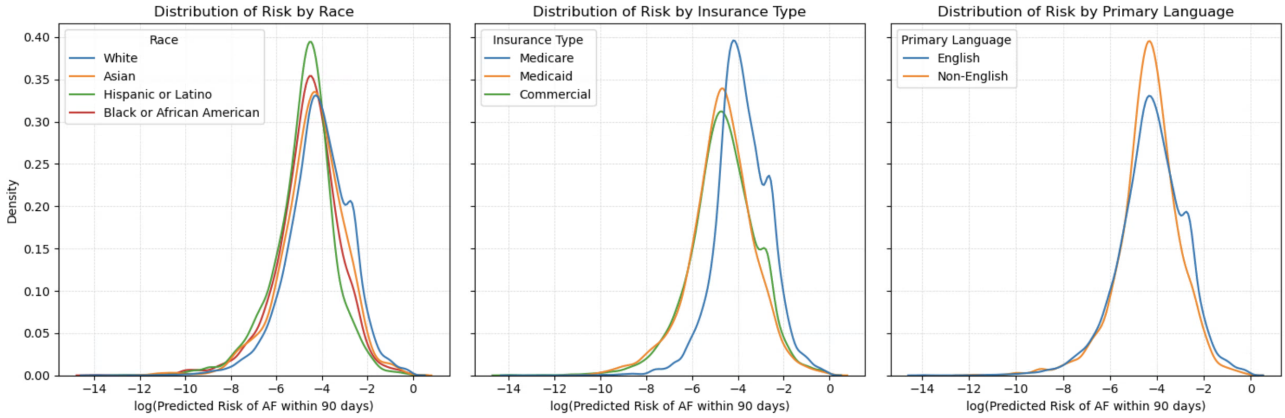

Figure 6: **Distributions of risk across study subgroups.** Each panel plots the distribution of risk among the demographic groups considered in our study, across race (left), insurance (middle), and primary spoken language (right).

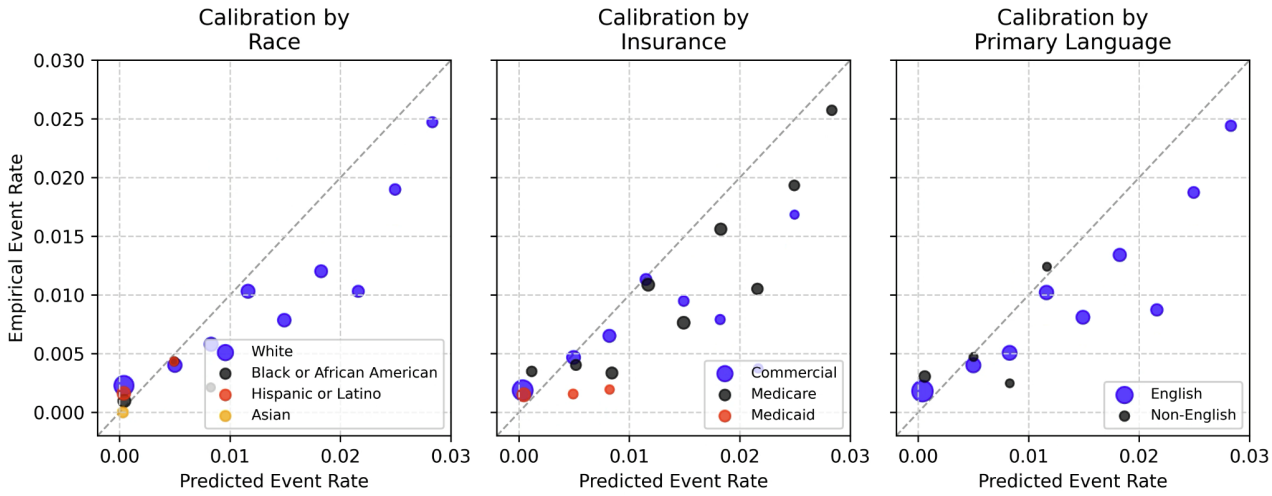

Figure 7: **Predicted event rates compared to empirical event rates.** We plot predicted event rates (x-axis) against empirical event rates (y-axis) in 5 bins evenly spaced between between 0 and 0.025. The diameter of each dot is proportional to the number of patients the dot represents and we exclude dots representing fewer than 20 patients. Figure 6 reports the full distribution of risk across subgroups.

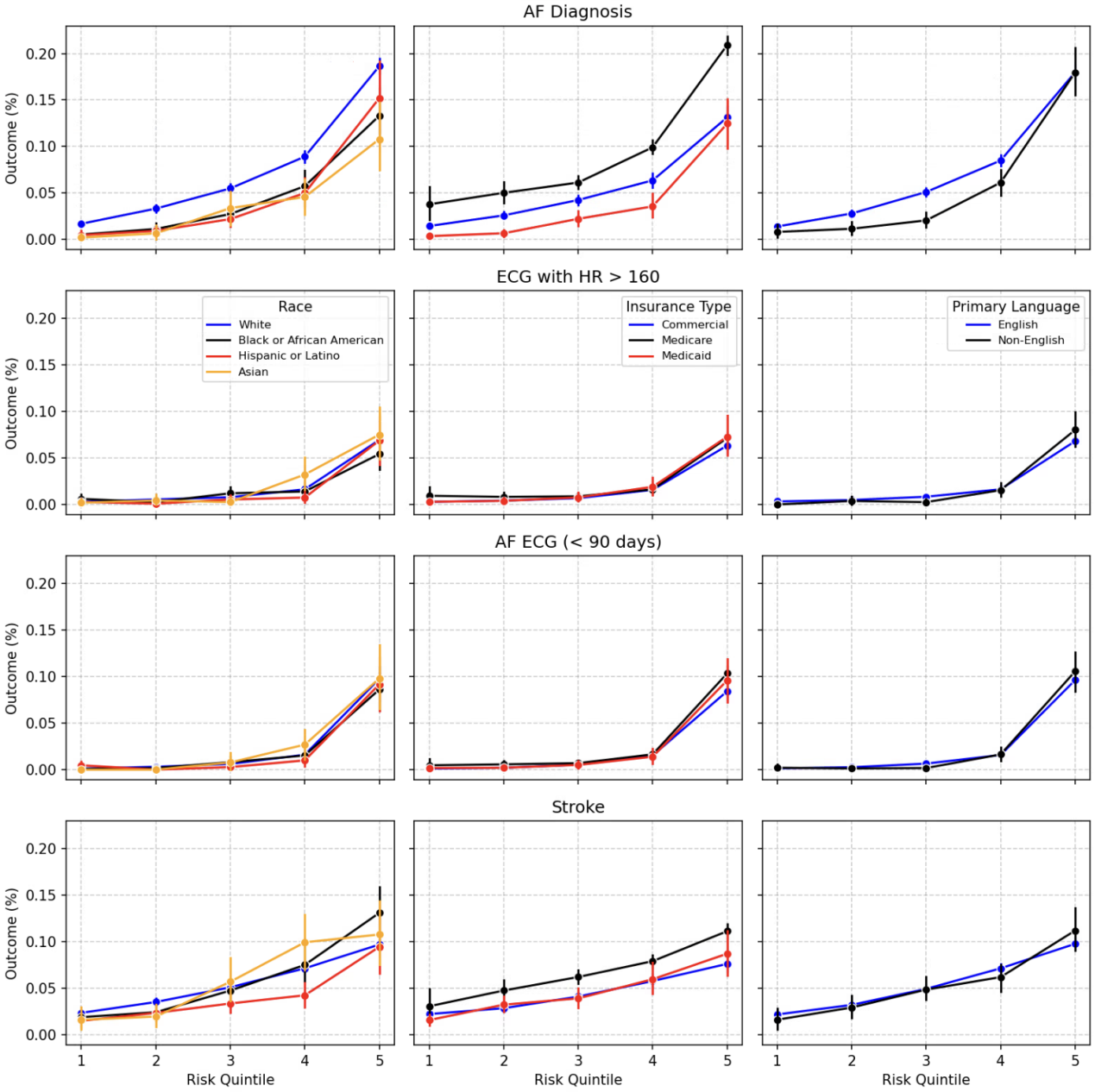

Figure 8: **Outcome rates across risk quintiles.** Here, we report outcome rates conditional on risk quintile across four outcomes: AF diagnosis (first row), an ECG with heart rate exceeding 160 within one year (second row), an AF ECG within three months (third row), and stroke within one year (fourth row). Each column corresponds to different demographic groups: race, insurance, and primary spoken language. Differences in diagnosis rates conditional on risk are not explained by differences in rates of AF complications conditional on risk.

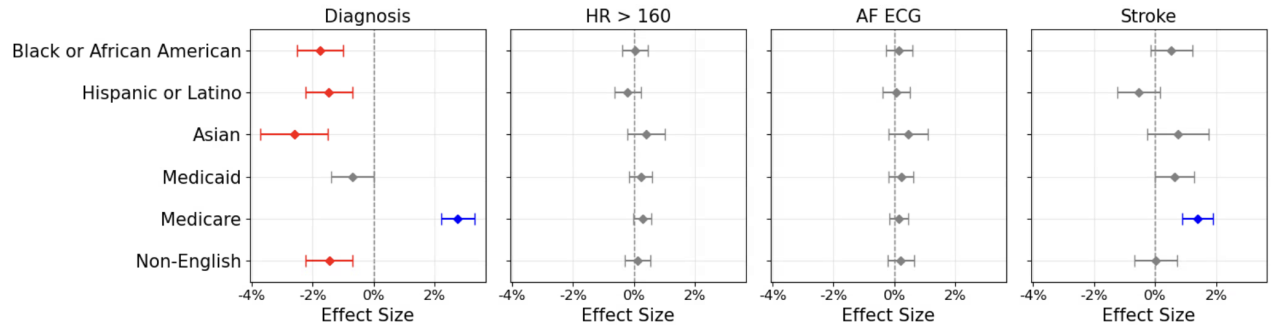

Figure 9: **Coefficients across all outcomes.** We expand on the effect sizes reported in the main text to include two additioanl AF complications: an AF ECG within 3 months, and stroke within one year. Trends are consistent, where there are no significant differences in rates of AF complications despite significant differences in diagnosis rates. The one except is the rate of stroke among Medicare patients; we attribute this behavior to how risk factors for stroke are more common among Medicare patients, relative to other groups.

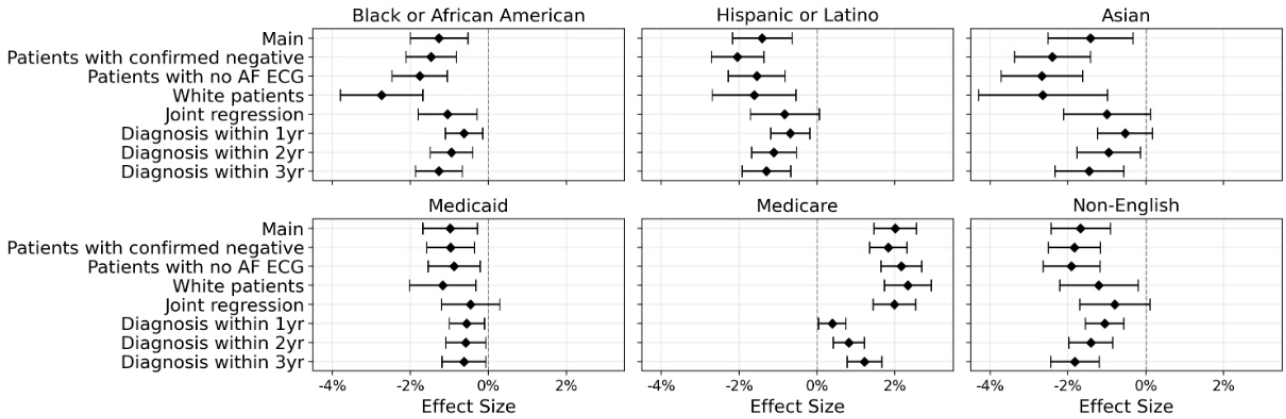

Figure 10: **Sensitivity analyses.** We perform several sensitivity analyses to assess the robustness of our results. The top row in each panel includes results from the main text for ease of comparison. We test multiple dataset construction criteria (white patients, patients with a confirmed normal ECG, patients with no downstream AF ECG), regression on an expanded set of controls (joint regression), and alternate definitions of diagnosis horizon (within 1 year, 2 years, and 3 years). Effect sizes are directionally consistent across sensitivity analyses. Effect sizes largely maintain significance, with the exception of effect sizes estimated for Medicaid patients and Non-English speakers, and diagnosis within a short time horizon.