

# **SUPPLEMENTAL MATERIAL**

**Table S1. Model specifications.**

Feature	Definition	Models
Kaiser	>1 outpatient ICD AF diagnosis billing code OR 1 outpatient ICD AF diagnosis billing code and ECG consistent with AF	Model 2
Medicare	≥1 inpatient ICD AF diagnosis billing code OR ≥2 outpatient ICD AF diagnoses billing codes within 365 days	Model 1
<i>Demographics: Based on status at the time of the index AF encounter</i>		
Age	Ordinal; roughly according to decade (18 to 29, 30 to 39, 40 to 49, 50 to 59, 60 to 69, 70 to 79, 80 and older)	Model 3 Model 5
Sex	Binary; female (reference) or male	Model 3 Model 5
Race	Categorical; white (reference), black, Asian, other/missing	Model 3 Model 5
Hispanic	Categorical; Hispanic, not Hispanic, or missing	Model 3 Model 5
Primary payer	Categorical; Medicare (reference), Medicaid, private, self, other/missing	Model 3 Model 5
Inpatient AF diagnosis	Binary; presence of an inpatient billing code at any time during the study period	Model 1 Model 3 Model 5
Number of outpatient AF diagnoses	Numeric; the total number of outpatient AF billing codes during the study period	Model 3 Model 5
Primary AF diagnosis	Binary; presence of an AF billing code in the first billing position any time during the study period	Model 3 Model 5
Acute myocardial infarction	Binary; presence of an ICD diagnosis billing code for condition any time during the study period	Model 3 Model 5
<i>Comorbid Conditions: Based on ICD billing codes</i>		
Coronary artery disease	Binary; presence of an ICD diagnosis billing code for condition any time during the study period	Model 3
Valvular heart disease		Model 5
Congestive heart failure		

Peripheral vascular disease		
Cerebrovascular disease		
Dementia		
Pulmonary heart disease		
Rheumatologic disease		
Gastrointestinal ulcer		
Liver disease		
Diabetes mellitus		
Acute renal disease		
Chronic renal disease		
Lymphoma		
Hypertension		
Coagulopathy		
Electrolyte disorder		
Anemia		
Cancer		
Dialysis		
<i>Procedures: Based on ICD billing codes</i>		
Heart valve surgery	Binary; presence of an ICD procedure billing code for procedure any time during the study period	Model 3
Coronary artery bypass grafting		Model 5
Percutaneous coronary intervention		
Angiogram		
Pacemaker/defibrillator		
Cardioversion (ICD based)		
<i>Procedures: Based on CPT billing codes</i>		
Cardioversion (CPT based)	Binary; presence of a CPT billing code for procedure any time during the study period	Model 5
Ablation		
<i>Electrocardiograms and text</i>		
Electrocardiogram	Binary; presence of an ECG interpretation that includes an AF-specific term any time during the study period	Model 2 Model 5
Number of AF mentions in the text	Categorical, split into zero and quartiles for values >0; the total number of AF mentions, as extracted by NLP, in the available text any time during the study period	Model 5
Non-negated AF mention in the text	Binary; the presence of a non-negated reference to AF in the available text any time during the study period	Model 4

Model Definitions (also see Methods section of manuscript):

Model #1: Outpatient and inpatient AF billing codes (Medicare methodology)<sup>2</sup>:  $\geq 1$  inpatient diagnosis or  $\geq 2$  outpatient diagnoses within 365 days

Model #2: Outpatient AF codes and electrocardiogram (Kaiser methodology)<sup>4</sup>:  $>1$  outpatient ICD code, 1 outpatient ICD code and ECG consistent with AF

Model #3: Demographics and International Classification of Diseases (ICD) billing AF codes: A logistic regression model using patient demographics, presence of an inpatient AF diagnosis, presence of a primary AF diagnosis, number of outpatient AF diagnoses, comorbid conditions and procedures from ICD codes, year of index AF diagnosis

Model #4: Natural language processing: At least one non-negated mention of AF in the clinical text

Model #5: Comprehensive: A comprehensive logistic regression model combining patient demographics, presence of an inpatient AF diagnosis, presence of a primary AF diagnosis, number of outpatient AF diagnoses, comorbid conditions and procedures from ICD codes, year of index AF diagnosis, at least one non-negated mention of AF in clinical text, ECG with reference to AF, CPT codes for ablation or cardioversion

AF=atrial fibrillation; CPT=current procedural terminology; ICD=International Classification of Diseases

**Table S2. Target terms used in natural language processing task to identify atrial fibrillation patients from clinical notes.**

Target term	Regular Expression
afib	\bafib\b \batrial\sfib a-fib a\.\sfib a\.fib \ba\sfib\b
<b>Modifier terms</b>	
no	\bno(?:\sfurther)\b
not	\bnot\b
none	\bnone\b
negative	\bnegative\b
denies	denies denied denying
family	\bmother\b \bfather\b \bsister\b \bbrother\b \bdaughter\b \bson\b \baunt\b \buncle\b \bgranddaughter\b \bgrandson\b
rule out	r/o r\o \brule\s+out\b \brules\s+out\b \bruled\s+out\b
unlikely	\bunlikely\b
investigate	\binvestigate\b \binvestigating\b
look for	\blook\s+for\b\b
differential	\bdifferential\b\b ddx
possible	\bpossible\b
holter	\b(holter event)\s+(monitor(ing)?\s+)?ordered\s+for\b\b ddx
etc	\betc\b
screen for	\bscreen\s+for\b
risk of	\brisk\s+(of for)\b
suspicious	\bsuspicious\b
question of	\bquestion\s+of\b

**Table S3. Training model regression results.**

KAISER MODEL						
Logistic regression	Number of obs	=	786			
	LR chi2(1)	=	78.77			
	Prob > chi2	=	0.0000			
Log likelihood = -349.45524	Pseudo R2	=	0.1013			
-----						
binary_adj_goldstd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
-----						
kaiser	1.723243	.2125683	8.11	0.000	1.306616	2.139869
_cons	.7256704	.110745	6.55	0.000	.5086141	.9427267
-----						
Logistic model for binary_adj_goldstd						
number of observations = 786						
area under ROC curve = 0.6951						
Empirical cutpoint estimation						
Method:		Liu				
Reference variable:		binary_adj_goldstd (0=neg, 1=pos)				
Classification variable:		kaiser_lr				
Empirical optimal cutpoint:		.7971682				
Sensitivity at cutpoint:		0.60				
Specificity at cutpoint:		0.79				
Area under ROC curve at cutpoint:		0.70				
(415 real changes made)						
Detailed report of sensitivity and specificity						
-----						
			Correctly			
Cutpoint	Sensitivity	Specificity	Classified	LR+	LR-	
-----						



Log likelihood = -356.13488 Pseudo R2 = 0.0841

---

binary_adj_goldstd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
simpleicd	1.675786	.2039045	8.22	0.000	1.27614	2.075431
_cons	.1563462	.1689453	0.93	0.355	-.1747804	.4874729

---

Logistic model for binary\_adj\_goldstd

number of observations = 786

area under ROC curve = 0.6509

Empirical cutpoint estimation

Method: Liu  
Reference variable: binary\_adj\_goldstd (0=neg, 1=pos)  
Classification variable: medicare\_lr  
Empirical optimal cutpoint: .70051131  
Sensitivity at cutpoint: 0.88  
Specificity at cutpoint: 0.42  
Area under ROC curve at cutpoint: 0.65  
(645 real changes made)

Detailed report of sensitivity and specificity

---

Cutpoint	Sensitivity	Specificity	Correctly		
			Classified	LR+	LR-
( >= 1 )	87.97%	42.21%	79.01%	1.5223	0.2849

---

Obs	ROC		-Asymptotic Normal--	
	Area	Std. Err.	[95% Conf. Interval]	
786	0.6509	0.0210	0.60978	0.69205



binary_adj	medicare_class		
_goldstd	Pos.	Neg.	Total
Abnormal	556	76	632
Normal	89	65	154
Total	645	141	786

True abnormal diagnosis defined as binary\_adj\_goldstd = 1

[95% Confidence Interval]

Prevalence	Pr (A)	80%	77%	83.1%
Sensitivity	Pr (+ A)	88%	85.2%	90.4%
Specificity	Pr (- N)	42.2%	34.3%	50.4%
ROC area	(Sens. + Spec.)/2	.651	.61	.692
Likelihood ratio (+)	Pr (+ A)/Pr (+ N)	1.52	1.33	1.75
Likelihood ratio (-)	Pr (- A)/Pr (- N)	.285	.215	.377
Odds ratio	LR(+)/LR(-)	5.34	3.59	7.96
Positive predictive value	Pr (A +)	86.2%	83.3%	88.8%
Negative predictive value	Pr (N -)	46.1%	37.7%	54.7%

### ICD MODEL

Logistic regression	Number of obs	=	786
	LR chi2(43)	=	260.73
	Prob > chi2	=	0.0000
Log likelihood = -258.47662	Pseudo R2	=	0.3353

binary_adj_goldstd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
agegrp1	.0446484	.0099897	4.47	0.000	.025069	.0642279
sex	.358715	.2435289	1.47	0.141	-.1185929	.8360229

race_categ							
Black	-1.942464	1.308129	-1.48	0.138	-4.506349	.6214212	
Asian	-1.375892	1.244579	-1.11	0.269	-3.815221	1.063437	
Other/missing	-.4826733	.4840383	-1.00	0.319	-1.431371	.4660243	
hispanic	.0927108	.1684888	0.55	0.582	-.2375211	.4229427	
index_payl_categ							
2	.8464419	.6084834	1.39	0.164	-.3461637	2.039047	
3	.0162844	.3308754	0.05	0.961	-.6322194	.6647882	
4	.4823887	.6985767	0.69	0.490	-.8867964	1.851574	
5	2.056091	1.095191	1.88	0.060	-.0904435	4.202626	
inpatientdx	1.640814	.3087453	5.31	0.000	1.035684	2.245943	
countoutpatient_afib	.4895864	.0782739	6.25	0.000	.3361723	.6430004	
afibicd_primary	-.9130516	.2387842	-3.82	0.000	-1.38106	-.4450431	
index_year	.0646928	.0550775	1.17	0.240	-.0432571	.1726428	
amidiag_all	.3809207	.4520047	0.84	0.399	-.5049923	1.266834	
caddiag_all	.0267004	.2832404	0.09	0.925	-.5284405	.5818414	
valveddiag_all	-.121885	.2692074	-0.45	0.651	-.6495218	.4057518	
chfdiag_all	.2366036	.3143618	0.75	0.452	-.3795342	.8527414	
pvddiag_all	-.6911086	.3083174	-2.24	0.025	-1.2954	-.0868175	
cvddiag_all	-1.278336	.2852031	-4.48	0.000	-1.837324	-.7193479	
dementiadiag_all	-.0838705	.3846242	-0.22	0.827	-.8377201	.669979	
pulmdzdiag_all	-.2523725	.2855902	-0.88	0.377	-.812119	.3073739	
rheumdiag_all	.7370864	.5824089	1.27	0.206	-.404414	1.878587	
ulcerdiag_all	.0170389	.5962036	0.03	0.977	-1.151499	1.185576	
liverdiag_all	-.4056906	.3076358	-1.32	0.187	-1.008646	.1972644	
dmdiag_all	-.0513154	.2568892	-0.20	0.842	-.554809	.4521781	
renaldiag_all	.0839594	.4110108	0.20	0.838	-.7216069	.8895257	
ckddiag_all	.7758547	.4285092	1.81	0.070	-.0640079	1.615717	
lymphdiag_all	.7620735	.8795399	0.87	0.386	-.961793	2.48594	
pulmhtndiag_all	.1058361	.3386697	0.31	0.755	-.5579443	.7696164	
htndiag_all	.0271604	.2786255	0.10	0.922	-.5189355	.5732563	

thyroiddiag_all		.2414638	.2949052	0.82	0.413	-.3365399	.8194674
coagdiag_all		.3509488	.3714666	0.94	0.345	-.3771123	1.07901
elecdiag_all		-.1480849	.3050795	-0.49	0.627	-.7460297	.44986
anemiadiag_all		.1875522	.3232971	0.58	0.562	-.4460984	.8212028
cancerdiag_all		.3973601	.2820185	1.41	0.159	-.155386	.9501062
dialysis_icdproc_all		-.3362597	.7277529	-0.46	0.644	-1.762629	1.09011
valve_icdproc_all		-.1382759	.6785085	-0.20	0.839	-1.468128	1.191576
cabg_icdproc_all		-.0303663	.8015717	-0.04	0.970	-1.601418	1.540685
pci_icdproc_all		.2712517	.9442262	0.29	0.774	-1.579398	2.121901
angio_icdproc_all		.0102707	.6437773	0.02	0.987	-1.25151	1.272051
ppm_defib_icdproc_all		-.8168417	.6898823	-1.18	0.236	-2.168986	.5353027
dccv_icdproc_all		1.196939	.8358149	1.43	0.152	-.4412278	2.835106
_cons		-132.932	110.9735	-1.20	0.231	-350.436	84.57198

-----  
Note: 0 failures and 44 successes completely determined.

Logistic model for binary\_adj\_goldstd

number of observations = 786

area under ROC curve = 0.8738

Empirical cutpoint estimation

Method: Liu  
Reference variable: binary\_adj\_goldstd (0=neg, 1=pos)  
Classification variable: icd\_lr  
Empirical optimal cutpoint: .77046734  
Sensitivity at cutpoint: 0.78  
Specificity at cutpoint: 0.81  
Area under ROC curve at cutpoint: 0.80

(525 real changes made)

Detailed report of sensitivity and specificity

-----  
Correctly  
Cutpoint      Sensitivity      Specificity      Classified      LR+      LR-  
-----

( >= 1 )              78.48%              81.17%              79.01%              4.1676              0.2651

-----

```

                ROC                -Asymptotic Normal--
Obs         Area         Std. Err.         [95% Conf. Interval]
-----
786         0.7982         0.0178         0.76337         0.83313

```

binary_adj	icd_class		
_goldstd	Pos.	Neg.	Total
Abnormal	496	136	632
Normal	29	125	154
Total	525	261	786

True abnormal diagnosis defined as binary\_adj\_goldstd = 1

[95% Confidence Interval]

Prevalence                              Pr (A)              80%              77%              83.1%

Sensitivity                              Pr (+|A)              78.5%              75.1%              81.6%

Specificity                              Pr (-|N)              81.2%              74.1%              87%

ROC area                              (Sens. + Spec.)/2              .798              .763              .833

Likelihood ratio (+)              Pr (+|A)/Pr (+|N)              4.17              2.99              5.8

Likelihood ratio (-)              Pr (-|A)/Pr (-|N)              .265              .224              .313

Odds ratio                              LR(+)/LR(-)              15.7              10.1              24.5

Positive predictive value              Pr (A|+)              94.5%              92.2%              96.3%

Negative predictive value              Pr (N|-)              47.9%              41.7%              54.1%

### NLP MODEL

Logistic regression    Number of obs              =              786

```

LR chi2(1) = 287.78
Prob > chi2 = 0.0000
Pseudo R2 = 0.3701
Log likelihood = -244.94682

```

```

-----
binary_adj_goldstd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
afnlp_mrn_predict |  4.061283   .2916117   13.93   0.000   3.489735   4.632831
      _cons | -1.684339   .2566414   -6.56   0.000  -2.187347  -1.181331
-----

```

Logistic model for binary\_adj\_goldstd

number of observations = 786

area under ROC curve = 0.8007

Empirical cutpoint estimation

```

Method:                               Liu
Reference variable:                   binary_adj_goldstd (0=neg, 1=pos)
Classification variable:              afnlp_lr
Empirical optimal cutpoint:           .53578696
Sensitivity at cutpoint:              0.97
Specificity at cutpoint:              0.63
Area under ROC curve at cutpoint:     0.80
(671 real changes made)

```

Detailed report of sensitivity and specificity

```

-----
                          Correctly
Cutpoint   Sensitivity   Specificity   Classified   LR+   LR-
-----
( >= 1 )   97.15%         62.99%       90.46%       2.6248  0.0452
-----

```

```

ROC
Obs   Area   Std. Err.   [95% Conf. Interval]
-----
-Asymptotic Normal--
-----

```

786      0.8007      0.0198      0.76189      0.83949

binary_adj	afnlp_class		
_goldstd	Pos.	Neg.	Total
Abnormal	614	18	632
Normal	57	97	154
Total	671	115	786

True abnormal diagnosis defined as binary\_adj\_goldstd = 1

[95% Confidence Interval]

Prevalence	Pr (A)	80%	77%	83.1%
Sensitivity	Pr (+ A)	97.2%	95.5%	98.3%
Specificity	Pr (- N)	63%	54.8%	70.6%
ROC area	(Sens. + Spec.)/2	.801	.762	.839
Likelihood ratio (+)	Pr (+ A)/Pr (+ N)	2.62	2.14	3.23
Likelihood ratio (-)	Pr (- A)/Pr (- N)	.0452	.0282	.0724
Odds ratio	LR (+)/LR (-)	58	32.9	102
Positive predictive value	Pr (A +)	91.5%	89.1%	93.5%
Negative predictive value	Pr (N -)	84.3%	76.4%	90.5%

### COMPREHENSIVE MODEL

Logistic regression	Number of obs	=	786
	LR chi2(47)	=	459.16
	Prob > chi2	=	0.0000
Log likelihood = -159.26033	Pseudo R2	=	0.5904

binary_adj_goldstd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
--------------------	-------	-----------	---	------	----------------------

afnlp_mrn_predict		2.2063	.5049817	4.37	0.000	1.216554	3.196046
agegrp1		.0328202	.0130682	2.51	0.012	.0072069	.0584335
sex		.006181	.3288827	0.02	0.985	-.6384173	.6507793
race_categ							
Black		-2.187707	1.514649	-1.44	0.149	-5.156364	.7809499
Asian		.2546778	3.239757	0.08	0.937	-6.095129	6.604485
Other/missing		-.656952	.6036018	-1.09	0.276	-1.83999	.5260858
hispanic		.1670156	.2225115	0.75	0.453	-.269099	.6031302
index_pay1_categ							
2		.5163838	.7883102	0.66	0.512	-1.028676	2.061443
3		-.2421059	.4610076	-0.53	0.599	-1.145664	.6614523
4		.9414504	.8906946	1.06	0.291	-.804279	2.68718
5		2.136887	1.231747	1.73	0.083	-.2772917	4.551066
inpatientdx		-.1369238	.4172727	-0.33	0.743	-.9547633	.6809157
countoutpatient_afib		.1403936	.0740849	1.90	0.058	-.0048101	.2855974
afibicd_primary		-1.262697	.3451367	-3.66	0.000	-1.939153	-.5862419
index_year		-.0405696	.075878	-0.53	0.593	-.1892877	.1081485
amidiag_all		.561467	.6321116	0.89	0.374	-.6774491	1.800383
caddiag_all		-.0842321	.3797421	-0.22	0.824	-.8285129	.6600487
valveddiag_all		-.0823799	.3639308	-0.23	0.821	-.7956711	.6309113
chfdiag_all		-.113876	.4480115	-0.25	0.799	-.9919625	.7642104
pvddiag_all		-.6556172	.4073686	-1.61	0.108	-1.454045	.1428106
cvddiag_all		-1.12224	.3907799	-2.87	0.004	-1.888155	-.3563258
dementiadiag_all		.0200449	.5326683	0.04	0.970	-1.023966	1.064056
pulmdzdiag_all		-.3854925	.3913458	-0.99	0.325	-1.152516	.3815311
rheumdiag_all		.5394939	.8438259	0.64	0.523	-1.114374	2.193362
ulcerdiag_all		.2380188	.8312635	0.29	0.775	-1.391228	1.867265
liverdiag_all		-.4841053	.4199376	-1.15	0.249	-1.307168	.3389572
dmdiag_all		-.0962111	.3521609	-0.27	0.785	-.7864338	.5940115
renaldiag_all		-.2154304	.5936916	-0.36	0.717	-1.379045	.9481838

ckddiag_all		1.626106	.6297206	2.58	0.010	.391876	2.860335
lymphdiag_all		.5713589	1.101511	0.52	0.604	-1.587562	2.73028
pulmhtndiag_all		.4727443	.4709309	1.00	0.315	-.4502633	1.395752
htndiag_all		-.5871352	.3913627	-1.50	0.134	-1.354192	.1799215
thyroiddiag_all		.012911	.4038323	0.03	0.974	-.7785859	.8044078
coagdiag_all		.4569073	.5011965	0.91	0.362	-.5254199	1.439234
elecdiag_all		-.613357	.4176761	-1.47	0.142	-1.431987	.2052732
anemiadiag_all		.2884396	.4747701	0.61	0.543	-.6420927	1.218972
cancerdiag_all		.1483555	.38792	0.38	0.702	-.6119538	.9086648
dialysis_icdproc_all		-.9619088	1.086018	-0.89	0.376	-3.090466	1.166648
valve_icdproc_all		-1.106743	.9657247	-1.15	0.252	-2.999529	.7860425
cabg_icdproc_all		1.961944	1.187286	1.65	0.098	-.3650924	4.288981
pai_icdproc_all		-.623917	1.42916	-0.44	0.662	-3.425019	2.177185
angio_icdproc_all		-.3559314	.8904829	-0.40	0.689	-2.101246	1.389383
ppm_defib_icdproc_all		.0118631	1.02624	0.01	0.991	-1.99953	2.023257
dccv_icdcpt_binary		-.0867684	1.027261	-0.08	0.933	-2.100162	1.926625
ablate_cpt_binary		-1.658924	1.220243	-1.36	0.174	-4.050557	.7327092
ecg_afib		2.881581	.6455071	4.46	0.000	1.616411	4.146752
q_afnlp_total		1.182809	.2549187	4.64	0.000	.683178	1.682441
_cons		78.62454	152.8337	0.51	0.607	-220.9241	378.1731

-----  
Note: 0 failures and 14 successes completely determined.

Logistic model for binary\_adj\_goldstd

number of observations = 786

area under ROC curve = 0.9504

Empirical cutpoint estimation

Method: Liu  
Reference variable: binary\_adj\_goldstd (0=neg, 1=pos)  
Classification variable: comprehensive\_lr  
Empirical optimal cutpoint: .7892637  
Sensitivity at cutpoint: 0.90  
Specificity at cutpoint: 0.87  
Area under ROC curve at cutpoint: 0.89



(591 real changes made)

Detailed report of sensitivity and specificity

Cutpoint	Sensitivity	Specificity	Correctly Classified	LR+	LR-
( >= 1 )	90.35%	87.01%	89.69%	6.9568	0.1109

Obs	ROC Area	Std. Err.	-Asymptotic Normal-- [95% Conf. Interval]
786	0.8868	0.0148	0.85779 0.91582

binary_adj   comprehensive_class			
_goldstd	Pos.	Neg.	Total
Abnormal	571	61	632
Normal	20	134	154
Total	591	195	786

True abnormal diagnosis defined as binary\_adj\_goldstd = 1

		[95% Confidence Interval]		
Prevalence	Pr(A)	80%	77%	83.1%
Sensitivity	Pr(+ A)	90.3%	87.8%	92.5%
Specificity	Pr(- N)	87%	80.7%	91.9%
ROC area	(Sens. + Spec.)/2	.887	.858	.916
Likelihood ratio (+)	Pr(+ A)/Pr(+ N)	6.96	4.62	10.5
Likelihood ratio (-)	Pr(- A)/Pr(- N)	.111	.0867	.142
Odds ratio	LR(+)/LR(-)	62.7	36.7	107

Positive predictive value	$\Pr(A +)$	96.6%	94.8%	97.9%
Negative predictive value	$\Pr(N -)$	68.7%	61.7%	75.2%

-----