ARIC Manuscript Proposal #3845

PC Reviewed: 5/11/21	**Status: _____**	**Priority: 2**
SC Reviewed: _________	**Status: _____**	**Priority: ____**

1.a. Full Title: *Performance of phenotyping algorithms in a cohort data set of validated events: The Atherosclerosis Risk in Communities Study*

b. Abbreviated Title (Length 26 characters): **Phenotyping algorithms**

2. Writing Group:
Writing group members:

Bailey M. DeBarmore, Kellan E. Ashley, Sara B. Jones, Alexander P. Keil, Jennifer L. Lund, Stuart D. Russell, Brent A. Williams, Wayne D. Rosamond

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. _BD____ **[please confirm with your initials electronically or in writing]**

First author: **Bailey DeBarmore**
Address: 123 W. Franklin St Chapel Hill NC Suite 410

Phone: 919-757-2266 Fax:
E-mail: bdebarmo@live.unc.edu

ARIC author to be contacted if there are questions about the manuscript and the first author does not respond or cannot be located (this must be an ARIC investigator).

Name: **Wayne D. Rosamond**
Address: 123 W. Franklin St Chapel Hill NC Suite 410

Phone: (919) 962-3230 Fax:
E-mail: wayne_rosamond@unc.edu

3. Timeline: This paper will be Paper 2 of the first author's dissertation work. Work will begin upon obtaining approval with goal journal submission in November 2021.

4. Rationale:

Routinely collected electronic healthcare data is increasingly being used for chronic disease case finding for study enrollment and for secondary research analysis.[1–3] Electronic phenotyping algorithms can be rule-based or use machine learning or both and may include structured and unstructured data elements.[4] These algorithms may be used to identify individuals eligible for a study (case finding) or to classify exposure status, outcome status, or comorbidity status of individuals for secondary research. Because routinely collecting electronic healthcare data found in electronic health records (EHR) is not collected for the purpose of research, it is important to understand the limitations of the data quality and the implications on algorithm validity.[5–8] However, while many algorithms may be commonly used, many are not validated or, when validation is done, accuracy measures are not always reported.[9–12] Furthermore, the accuracy of an algorithm validated in one study population may not generalize to a different study population.[13,14]

Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) are used as measures of accuracy to quantify the degree of misclassification. Prioritizing one of these accuracy measures over another depends on the research question at hand and on whether the algorithm is being used to ascertain exposure status, outcome status, or potential confounders such as comorbidities. An *outcome* classification algorithm with perfect specificity, even with low sensitivity, will result in an unbiased *relative* measure of effect, assuming nondifferential misclassification with respect to the exposure.[14] In contrast, selecting an *exposure* classification algorithm with high sensitivity is important particularly when the exposure is common.[14] High sensitivity classification algorithms are useful as an initial screen to pare down the potential study population prior to a more accurate but costly measurement tool.[14] For example, researchers planning to conduct manual chart review to identify acute myocardial infarction (MI) patients may wish to reduce the time and cost spent abstracting information by first applying a highly sensitivity phenotyping algorithm. High sensitivity algorithms are also preferred when the researchers wish to identify every possible patient eligible for a research study,[15] particularly if further eligibility will be confirmed at a later point, such as through individual phone interviews. Finally, some low sensitivity algorithms may have differential sensitivity depending on disease severity, given that patients with more severe disease may have more data available.[16] Thus, it is important to use high sensitivity algorithms to capture a study population representative of the entire disease spectrum, or in other words, to improve generalizability.[14,15] Positive predictive value and NPV are related to prevalence, sensitivity, and specificity. When researchers are willing to miss some false negatives for the benefit of ensuring those included truly have the condition of interest (true positives) it would be best to select an algorithm with both high specificity and high PPV.[14] When researchers wish to exclude individuals with a certain condition (and thus want to be sure those included are true negatives), it would be important to select an algorithm for that condition with high NPV.[14] Given some misclassification, it is important to weigh the benefits and costs of high sensitivity or high specificity against the goal of a research question.

We chose to focus on acute myocardial infarction (AMI) and heart failure (HF) in these analyses because patients with these conditions present differently to care. There is a lack of papers describing validation of electronic phenotyping algorithms using ICD-10-CM codes for these two conditions in the US. Validation of commonly used electronic phenotyping algorithms for AMI and HF, applied to EHR data, are needed to verify accuracy and assess misclassification.[3,5,13,17–22] We sought to fill this gap by calculating validation measures for several electronic phenotyping algorithms for AMI and HF in a US cohort study with both EHR data and event classification via physician review.

5. Main Hypothesis/Study Questions:

How do rule-based electronic phenotyping algorithms perform against physician-ascertained event classification for acute myocardial infarction and heart failure?

6. Design and analysis (study design, inclusion/exclusion, outcome and other variables of interest with specific reference to the time of their collection, summary of data analysis, and any anticipated methodologic limitations or challenges if present).

Analytic Sample: The ARIC cohort surveillance datasets will be left truncated at October 1, 2015 to correspond to the ICD-10-CM era. The data will be right-censored at the latest surveillance update (likely the end of ARIC 2019 cohort surveillance, including updated death data). Data from visits will be used to collect comorbidity information.

Gold Standard Classification: The MI or HF hospitalization considered the qualifying event will be identified in the ARIC event file using the MMCC physician's preferred diagnosis for MI or final HF MMCC classification. Event dates will be cross-checked between the most up to date ARIC events file and the surveillance data sets to gather information from the correct hospitalization.

Phenotyping Algorithms: Algorithm 2 and Algorithm 3 (both A and B) will be evaluated in the ARIC cohort event surveillance dataset (Table 1). Table 2 (MI) and Table 3 (HF) list the variable names and corresponding ARIC datasets that will be used to construct each phenotyping algorithm listed in Table 1. These tables also include the variables corresponding to the underlying components (e.g. ECG, pain, and biomarker evidence for MI diagnostic algorithm) of each ARIC classification.

Table 1. Phenotyping algorithms for evaluation in the ARIC cohort event surveillance data

	Acute Myocardial Infarction	**Hospitalized Heart Failure**
Algorithm 2A	(I21 or I22) in any position in hospital discharge list	(I50, I13.0, I13.2, or I11.0) in any position in hospital discharge list
Algorithm 2B	(I21 or I22) in primary or secondary position in hospital discharge list	(I50, I13.0, I13.2, or I11.0) in primary or secondary position in hospital discharge list
Algorithm 3A	(I21 or I22) in any position in hospital discharge list *AND* Elevated cardiac biomarker (troponin I, troponin T, CK-MB) *OR* cardiac procedure during hospitalization	(I50, I13.0, I13.2, or I11.0) in any position in hospital discharge list *AND* inpatient administration of IV diuretics *OR* (elevated BNP >500 pg/mL or elevated NT-proBNP >900 pg/mL for
Algorithm 3B	(I21 or I22) in primary or secondary position in hospital discharge list *AND* Elevated cardiac biomarker (troponin I, troponin T, CK-MB) *OR* cardiac procedure during hospitalization	(I50, I13.0, I13.2, or I11.0) in primary or secondary position in hospital discharge list *AND* inpatient administration of IV diuretics *OR* (elevated BNP >500 pg/mL or elevated NT-proBNP >900 pg/mL)

For acute MI, the ICD-10-CM codes of interest will be I21 (Acute myocardial infarction) and I22 (Subsequent myocardial infarction). These codes have been used in previous studies identifying myocardial infarction hospitalizations.[23–27] ICD-10-CM codes I21 and I22 include cardiac infarction, coronary embolism, occlusion, rupture, and thrombosis; and heart, myocardium, or ventricle infarction. I22 also includes recurrent myocardial infarction; myocardium reinfarction; heart, myocardium, or ventricle rupture, and subsequent type 1 myocardial infarction. Subsequent myocardial infarctions are those occurring within four weeks, or 28 days, of a previous acute myocardial infarction. In epidemiologic analyses of cohort studies, such as the ARIC Study, multiple MIs occurring within 28 days are typically considered to be the same event.

For HF, the ICD-10-CM codes of interest will be I50 (Heart failure), I13.0 and I13.2 (Hypertensive heart disease and chronic kidney disease with heart failure), and I11.0 (Hypertensive heart disease with heart failure). Medicare-based EHR HF studies utilized ICD-9-CM codes that map to these 4 ICD-10-CM codes. Researchers using the Clinical Practice Research Datalink in the United Kingdom have used a broader inclusive HF algorithm that also included ICD-10 codes for pulmonary embolism, pericarditis, cardiomyopathy, and rheumatic HF.[28,29]

Phenotypic Comparisons

The sample captured by each algorithm-ARIC classification subgroup (or case definition) will be compared on key phenotypic variables, such as demographics (including education level), comorbidities, and disease severity. Algorithm-classification subgroups refer to the permutations of Algorithm 2A – Definite MI, Algorithm 2A – Algorithm 2B…Algorithm 3B – Definite MI, and so on for all ARIC classifications, and again for HF. Dummy tables for these tabulations are shown in Table 4 (MI) and Table 5 (HF) at the end of this document. I will make phenotypic comparisons for all ARIC classifications separately (definite, probable, suspect, no MI and HF categories A through E) as well as commonly used groupings (definite/probable, suspect/no MI, and A/B, C, D/E). Where possible, the same variables tabulated from the EHR in Aims 1 and 2 are included from ARIC documentation, though these variables are not defined via diagnostic codes in ARIC. **Table 6** (MI) and **Table 7** (HF) at the end of this document list the variable names and corresponding ARIC datasets that will be used for the independent measures used to compare each algorithm-classification subgroup.

Additional Tabulations

ARIC surveillance data includes full medical record abstraction (structured and unstructured data). I will take advantage of the additional depth and breadth of data available to further describe the populations captured by each phenotyping algorithm.

The ARIC MI algorithm utilizes chest pain symptoms, cardiac biomarker evidence, and electrocardiogram evidence. The proportion of MI cases identified by each phenotyping algorithm meeting the varying levels of evidence for each of these data points will be presented in a table like Table 8.

Table 9 distinguishes between characteristics determined via transthoracic echocardiogram and transesophageal echocardiogram, such as dilated left ventricle, dilated right ventricle, impaired left ventricle systolic function, and impaired right ventricle systolic function. Characteristics determined from either echocardiogram method will be presented for publication combined, with a “yes” from either method qualifying.

Variable Definitions
This section defines the specific variables corresponding to measures that I will use to describe populations captured in each algorithm. There are additional tables in the Appendix describing the variable names and corresponding datasets. Some measures have multiple corresponding variables from different data sources collected in the ARIC study, such as history of diabetes recorded at the MI hospitalization versus measured at the most recent visit. Values from multiple sources for a single item will be compared.

Demographics: Age at the time of hospitalization will be calculated using the event date and date of birth. Gender, race, sex, and center will be crosschecked between the hospitalization dataset and visit 7 dataset as a quality control measure. The minimum age of the analytic population is expected to be 74, as that is the youngest age possible among ARIC participants in 2016.

Body Mass Index: For hospitalized MI, body mass index is not extracted from medical records. I will tabulate body mass index recorded at visit 7 by algorithm-classification group. For hospitalized HF, body mass index at discharge is extracted from hospitalization event medical records. Mean (SD) body mass index by algorithm group as well as categorized body mass index will be tabulated. I will also compare body mass index from the hospitalization to documented body mass index at visit 7.

Smoking Status: For hospitalized MI, smoking status as reported during the event is extracted from the medical records and will be tabulated in addition to smoking status recorded at visit 7. For hospitalized HF, smoking status from visit 7 will be tabulated. Smoking status recorded at visits is provided in several binary variables (current smoker (yes/no), former smoker (yes/no), ever smoker (yes/no)) and as a categorical variable (current, former, ever smoker).

Hypertension: For both hospitalized MI and HF, history of hypertension is recorded at the time of hospitalization and extracted from the medical record. Hypertension will also be defined using visit 7 data (SBP ≥ 140 or DBP ≥ 90 or self-report/catalogued use of anti-hypertensive medications). Note that catalogued use of medication refers to that ARIC participants are asked to bring all medication prescription bottles to ARIC visits for review and documentation by study staff.

Diabetes: For both hospitalized MI and HF, history of diabetes is recorded at the time of hospitalization and extracted from the medical record. Diabetes will also be defined using visit 7 data (fasting blood glucose ≥ 126 mg/dL or non-fasting blood glucose ≥ 200 or self-report/catalogued use of glucose-lowering medication).

Kidney Disease and Kidney Failure: The ARIC Study has 2 definitions for incident chronic kidney disease stage 3 or greater. Definition 1 includes participants that develop an eGFR-Cr <60 mL/min/1.73 m^2 AND an eGFR-Cr decline from baseline visit of at least 25% as recorded at study visits. Definition 2 includes Definition 1 but also includes US Renal Data System

(USRDS)-identified end-stage kidney disease events and cohort participants with hospitalizations or deaths with kidney disease-related ICD-9-CM or ICD-10-CM codes in any position (**Table 10** at the end of this document). The ARIC Study definition for incident kidney failure captures persons with USRDS-identified end stage kidney disease, eGFR-Cr <15 mL/min/1.73 m^2 at a study visit, or a hospitalization or death with kidney failure-related ICD-9-CM or ICD-10-CM codes in any position (**Table 11**). Prevalent kidney failure is identified via USRDS registry identification or eGFR-Cr <15 mL/min/1.73 m^2 at a previous study visit. For this analysis, I will use visit data to identify prevalent kidney disease (eGFR-Cr < 60 mL/min/1.73 m^2) and kidney failure (eGFR-Cr < 15 mL/min/1.73 m^2) as well as data from the incident files. For heart failure hospitalizations, report of dialysis use at the time of hospitalization is extracted from the medical record and will also be reported.

Mortality: Death data from proxy report, obituary review, or linkage with the National Death Index is recorded in the ARIC status and incidence files. Discharge disposition (alive/dead) will be used to determine in-hospital mortality. Death date and event date for MI or HF will be used to calculate 28-day and 1-year mortality, and will be all-cause mortality.

Heart Failure among Participants with Hospitalized MI: Heart failure events ascertained via surveillance or from visit self-report are included in the ARIC incident data files. Variables for incident HF (hospitalization, self-report, or death due to HF among those without prevalent HF at visit 1) and incident hospitalized HF post-visit 5 (2011 – 2013, the first visit after 2005 when heart failure adjudication began) are provided along with the associated date event. Prevalent HF at visit 1 is also included in the incidence file. There is also a specific variable for incident HF following hospitalized MI, with missing values for participants who had prevalent HF and then experienced an MI.

Atrial Fibrillation: Incident atrial fibrillation and the self-report date (or last date of semi- or annual follow-up prior to the end of visit 7) is provided in the ARIC incidence file and will be used for hospitalized MI and HF events. HF hospitalizations also have history of atrial fibrillation or flutter extracted from the hospital record. This data source will also be used to tabulate atrial fibrillation prevalence by HF algorithm-classification group.

History of Stroke or TIA: History of stroke or TIA prior to the ARIC study and incident ischemic stroke or TIA are documented in several ARIC datasets. For hospitalized MI and HF, history of stroke in the medical record is extracted into the surveillance datasets. History of stroke or TIA reported at visit 1 is included in the incidence dataset and prevalent stroke by the end of visit 7 is included in the visit 7 dataset. The incidence dataset also has a variable for definite or probable incident ischemic stroke with the associated hospitalized stroke admission date that can be used along with the MI or HF event date to determine if the stroke occurred before the qualifying event.

Statistical Analyses

Sensitivity, Specificity, PPV, and NPV

Algorithms 2A, 2B, 3A, and 3B for MI and HF will be cross-tabulated with ARIC surveillance classifications for MI and HF in several ways. For MI and HF, *r x c* contingency tables for combined groupings (definite/probable, suspect/no MI; A/B, C, D/E) (Table 12 for MI and Table 13 for HF at the end of this document) and for each classification group separately will be created (Table 14 for MI and Table 15 for HF at the end of this document) where A refers to definite acute decompensated heart failure, B refers to probable acute decompensated heart failure, C refers to chronic stable heart failure, D refers to unlikely heart failure, and E refers to unclassifiable heart failure per the MMCC adjudication.

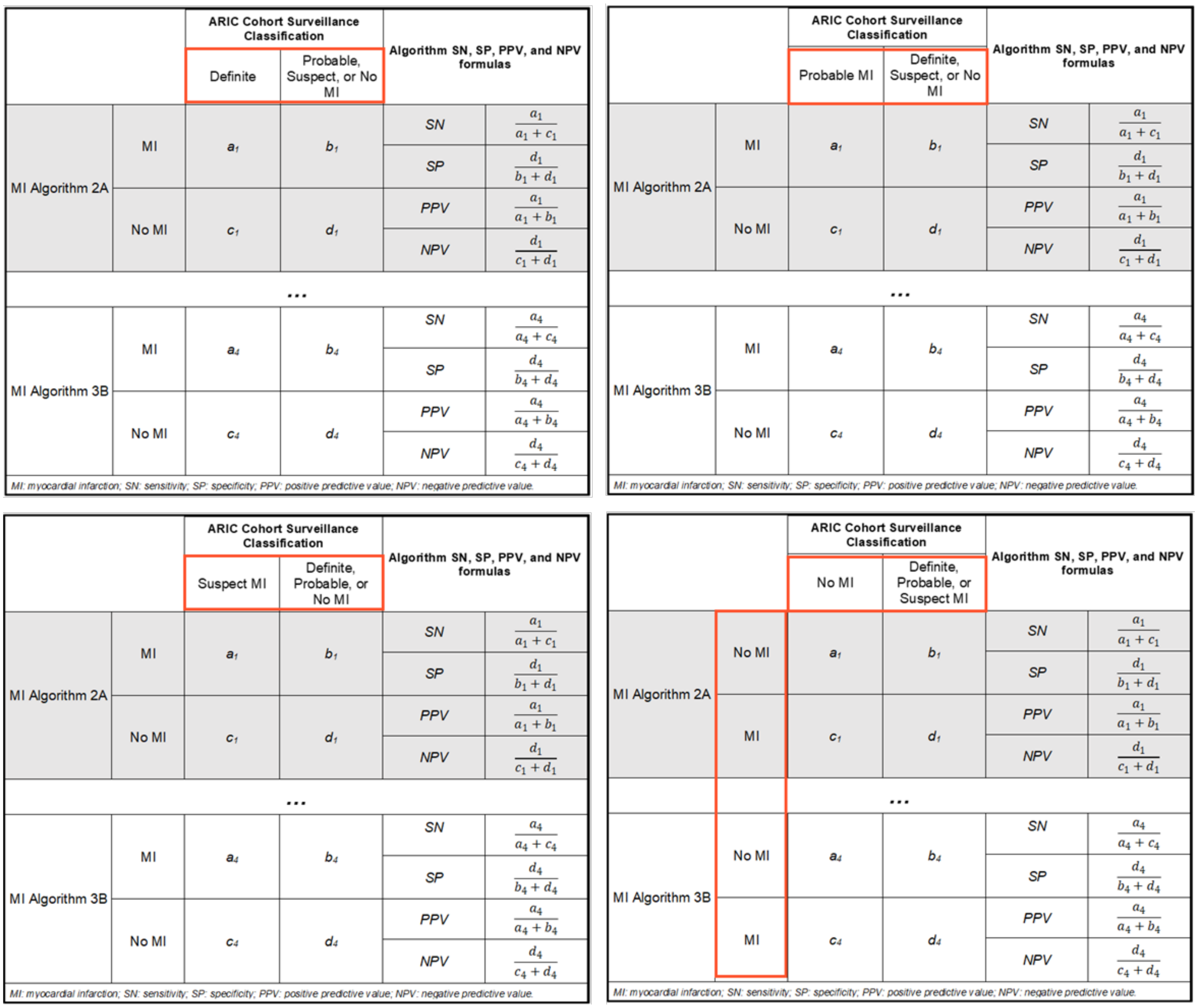
For the entire period of interest and separately by year, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) will be calculated for each algorithm against the separate classifications and the groupings, treating the ARIC classification as the gold standard. These values will be reported in a table like shell Table 16 for MI and shell Table 17 for HF. The tabulations by year are not shown in the dummy tables but I will replicate the tables for each year or alternatively include additional columns or rows to report values for each year in a single table. Changes over calendar time during the period of interest in sensitivity, specificity, PPV, and NPV will be evaluated visually.

For hospitalized MI in 2 categories, the calculations for sensitivity, specificity, PPV, and NPV are straight forward (Table 18 at the end of this document). Bootstrapped 95% confidence intervals for sensitivity, specificity, PPV, and NPV will be calculated using statistical software. For classifications with more than 2 categories, separate 2 x 2 tables for each algorithm and classification will be created to calculate sensitivity, specificity, PPV, NPV. For example, to calculate these measures for each MI classification, tables like those shown in Figure 1 on the next page will be constructed, treating each classification as a “positive” and all others as “negative”. The process for HF classifications grouped (A/B, C, D/E) and separate (A, B, C, D, E) will be the same. Bootstrapped 95% confidence intervals for sensitivity, specificity, PPV, and NPV will be calculated using statistical software.

Subgroups

Sensitivity, specificity, PPV, and NPV will also be calculated by age, race, and gender subgroups to determine if these measures for the MI algorithms and HF algorithms vary by population. The calculations described in the previous section will be repeated for age categories (74 – 84 years and 85 years and over), race groups (black and white), gender (men and women), and race-gender groups (white men, white women, black men, black women). Note that the youngest possible age among ARIC participants in 2016 is 74.

Subgroup sensitivity, specificity, PPV, and NPV will be presented in shell tables similar to those in the previous section and will be compared visually between subgroups using graphs. The implications for differing sensitivity, specificity, PPV, and NPV (if differences between subgroups are found) will be discussed with regard to interpreting research using EHR as secondary data, using algorithms to identify potential clinical trial populations, and the use of algorithms to estimate prevalence of cardiovascular disease at the national level.



		ARIC Cohort Surveillance Classification		Algorithm SN, SP, PPV, and NPV formulas	
		Definite	Probable, Suspect, or No MI		
MI Algorithm 2A	MI	a_1	b_1	*SN*	$\frac{a_1}{a_1 + c_1}$
				SP	$\frac{d_1}{b_1 + d_1}$
	No MI	c_1	d_1	*PPV*	$\frac{a_1}{a_1 + b_1}$
				NPV	$\frac{d_1}{c_1 + d_1}$
...					
MI Algorithm 3B	MI	a_4	b_4	*SN*	$\frac{a_4}{a_4 + c_4}$
				SP	$\frac{d_4}{b_4 + d_4}$
	No MI	c_4	d_4	*PPV*	$\frac{a_4}{a_4 + b_4}$
				NPV	$\frac{d_4}{c_4 + d_4}$

MI: myocardial infarction; SN: sensitivity; SP: specificity; PPV: positive predictive value; NPV: negative predictive value.

		ARIC Cohort Surveillance Classification		Algorithm SN, SP, PPV, and NPV formulas	
		Probable MI	Definite, Suspect, or No MI		
MI Algorithm 2A	MI	a_1	b_1	*SN*	$\frac{a_1}{a_1 + c_1}$
				SP	$\frac{d_1}{b_1 + d_1}$
	No MI	c_1	d_1	*PPV*	$\frac{a_1}{a_1 + b_1}$
				NPV	$\frac{d_1}{c_1 + d_1}$
...					
MI Algorithm 3B	MI	a_4	b_4	*SN*	$\frac{a_4}{a_4 + c_4}$
				SP	$\frac{d_4}{b_4 + d_4}$
	No MI	c_4	d_4	*PPV*	$\frac{a_4}{a_4 + b_4}$
				NPV	$\frac{d_4}{c_4 + d_4}$

MI: myocardial infarction; SN: sensitivity; SP: specificity; PPV: positive predictive value; NPV: negative predictive value.

		ARIC Cohort Surveillance Classification		Algorithm SN, SP, PPV, and NPV formulas	
		Suspect MI	Definite, Probable, or No MI		
MI Algorithm 2A	MI	a_1	b_1	*SN*	$\frac{a_1}{a_1 + c_1}$
				SP	$\frac{d_1}{b_1 + d_1}$
	No MI	c_1	d_1	*PPV*	$\frac{a_1}{a_1 + b_1}$
				NPV	$\frac{d_1}{c_1 + d_1}$
...					
MI Algorithm 3B	MI	a_4	b_4	*SN*	$\frac{a_4}{a_4 + c_4}$
				SP	$\frac{d_4}{b_4 + d_4}$
	No MI	c_4	d_4	*PPV*	$\frac{a_4}{a_4 + b_4}$
				NPV	$\frac{d_4}{c_4 + d_4}$

MI: myocardial infarction; SN: sensitivity; SP: specificity; PPV: positive predictive value; NPV: negative predictive value.

		ARIC Cohort Surveillance Classification		Algorithm SN, SP, PPV, and NPV formulas	
		No MI	Definite, Probable, or Suspect MI		
MI Algorithm 2A	No MI	a_1	b_1	*SN*	$\frac{a_1}{a_1 + c_1}$
				SP	$\frac{d_1}{b_1 + d_1}$
	MI	c_1	d_1	*PPV*	$\frac{a_1}{a_1 + b_1}$
				NPV	$\frac{d_1}{c_1 + d_1}$
...					
MI Algorithm 3B	No MI	a_4	b_4	*SN*	$\frac{a_4}{a_4 + c_4}$
				SP	$\frac{d_4}{b_4 + d_4}$
	MI	c_4	d_4	*PPV*	$\frac{a_4}{a_4 + b_4}$
				NPV	$\frac{d_4}{c_4 + d_4}$

MI: myocardial infarction; SN: sensitivity; SP: specificity; PPV: positive predictive value; NPV: negative predictive value.

Figure 1. Example 2 x 2 tables for each of the algorithm-subgroup classifications using 4 separate ARIC MI classifications (Algorithms 2B and 3A not shown)

7.a. Will the data be used for non-ARIC analysis or by a for-profit organization in this manuscript? ____ Yes _x___ **No**

b. If Yes, is the author aware that the current derived consent file ICTDER05 must be used to exclude persons with a value RES_OTH and/or RES_DNA = "ARIC only" and/or "Not for Profit" ? ____ Yes ____ No
(The file ICTDER has been distributed to ARIC PIs, and contains
the responses to consent updates related to stored sample use for research.)

8.a. Will the DNA data be used in this manuscript? ____ Yes _x___ **No**

8.b. If yes, is the author aware that either DNA data distributed by the Coordinating Center must be used, or the current derived consent file ICTDER05 must be used to exclude those with value RES_DNA = "No use/storage DNA"? ____ Yes ____ No

9. The lead author of this manuscript proposal has reviewed the list of existing ARIC Study manuscript proposals and has found no overlap between this proposal and previously approved manuscript proposals either published or still in active status. ARIC Investigators have access to the publications lists under the Study Members Area of the web site at: http://www.cscc.unc.edu/aric/mantrack/maintain/search/dtSearch.html

___x___ Yes ________ No

10. What are the most related manuscript proposals in ARIC (authors are encouraged to contact lead authors of these proposals for comments on the new proposal or collaboration)?

This paper will form Paper 2 for the first author's dissertation. Writing group member Jennifer Lund is part of the first author's dissertation committee, and has conducted previous research in ARIC using Medicare algorithms (MP 2542 Claims-based frailty in ARIC, 2015).

MP 3123 – Machine learning based phenotyping in heart failure (Sanchez Martinez and Soloman 2018) – This proposal aims to use echocardiographic data and unsupervised machine learning to phenotype heart failure patients. This proposal is similar in that it utilizes informatics-based methods but distinct enough from the aims of this paper that collaboration is not appropriate.

MP 3573 – AI-ECG for AF Prediction in ARIC (Noseworthy and Chen 2020) – This proposal aims to externally validate an AI-enabled ECG algorithm to identify patients with AF. This proposal is similar in that it utilizes informatics-based methods but distinct in that it focuses on ECG and AF.

MP 2734 – ML for MI Classification (Bogle and Heiss, 2016) – This proposal is related to the aims presented here in that it seeks to use informatics-based methods to classify acute MI but differs in that it is using machine learning rather than rule-based algorithms. It differs in that it proposed using community data rather than cohort data. To the first author's knowledge, the corresponding publication has not been published.

MP 3118 – Comparison of existing methods for algorithmic classification of dementia status (Gianattasio and Power 2018) – This proposal aims to use predictive algorithms to identify dementia with Visit 5 and 6 data. The methods are informatics-based but not related to cardiovascular disease or to the phenotyping algorithms proposed in this proposal.

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data? ____ Yes __x__ No

11.b. If yes, is the proposal

___ A. primarily the result of an ancillary study (list number* _________)

___ B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)* __________ __________ __________)

*ancillary studies are listed by number https://sites.cscc.unc.edu/aric/approved-ancillary-studies

12a. Manuscript preparation is expected to be completed in one to three years. If a manuscript is not submitted for ARIC review at the end of the 3-years from the date of the approval, the manuscript proposal will expire.

12b. The NIH instituted a Public Access Policy in April, 2008 which ensures that the public has access to the published results of NIH funded research. It is **your responsibility to upload manuscripts to PubMed Central** whenever the journal does not and be in compliance with this policy. Four files about the public access policy from http://publicaccess.nih.gov/ are posted in http://www.cscc.unc.edu/aric/index.php, under Publications, Policies & Forms. http://publicaccess.nih.gov/submit_process_journals.htm shows you which journals automatically upload articles to PubMed central.

REFERENCES

1. Roumia, M. & Steinhubl, S. Improving Cardiovascular Outcomes Using Electronic Health Records. *Curr. Cardiol. Rep.* **16**, 451 (2014).
2. Reimer, A. P., Milinovich, A. & Madigan, E. A. Data quality assessment framework to assess electronic medical record data for use in research. *Int. J. Med. Inf.* **90**, 40–47 (2016).
3. Hripcsak, G. & Albers, D. J. Next-generation phenotyping of electronic health records. *J. Am. Med. Inform. Assoc.* **20**, 117–121 (2013).
4. Nissen, F., Quint, J. K., Morales, D. R. & Douglas, I. J. How to validate a diagnosis recorded in electronic health records. *Breathe* **15**, 64–68 (2019).
5. Hoeven, L. R. van *et al.* Validation of multisource electronic health record data: an application to blood transfusion data. *BMC Med. Inform. Decis. Mak.* **17**, 107 (2017).
6. Brouwer, E. S. *et al.* Validation of Medicaid Claims-based Diagnosis of Myocardial Infarction Using an HIV Clinical Cohort: *Med. Care* **53**, e41–e48 (2015).
7. Stürmer, T. *et al.* Methodological considerations when analysing and interpreting real-world data. *Rheumatology* **59**, 14–25 (2020).
8. Davidson, J., Banerjee, A., Muzambi, R., Smeeth, L. & Warren-Gash, C. Validity of Acute Cardiovascular Outcome Diagnoses Recorded in European Electronic Health Records: A Systematic Review. *Clin. Epidemiol.* **Volume 12**, 1095–1111 (2020).
9. Rubbo, B. *et al.* Use of electronic health records to ascertain, validate and phenotype acute myocardial infarction: A systematic review and recommendations. *Int. J. Cardiol.* **187**, 705–711 (2015).

10. McCormick, N., Lacaille, D., Bhole, V. & Avina-Zubieta, J. A. Validity of Myocardial Infarction Diagnoses in Administrative Databases: A Systematic Review. *PLoS ONE* **9**, e92286 (2014).

11. McCormick, N., Bhole, V., Lacaille, D. & Avina-Zubieta, J. A. Validity of Diagnostic Codes for Acute Stroke in Administrative Databases: A Systematic Review. *PLOS ONE* **10**, e0135834 (2015).

12. McCormick, N., Lacaille, D., Bhole, V. & Avina-Zubieta, J. A. Validity of Heart Failure Diagnoses in Administrative Databases: A Systematic Review and Meta-Analysis. *PLoS ONE* **9**, e104519 (2014).

13. Manuel, D. G., Rosella, L. C. & Stukel, T. A. Importance of accurately identifying disease in studies using electronic health records. *BMJ* **341**, c4226–c4226 (2010).

14. Chubak, J., Pocobelli, G. & Weiss, N. S. Tradeoffs between accuracy measures for electronic health care data algorithms. *J. Clin. Epidemiol.* **65**, 343-349.e2 (2012).

15. Yao, R. J. R. *et al.* Sensitivity, specificity, positive and negative predictive values of identifying atrial fibrillation using administrative data: a systematic review and meta-analysis. *Clin. Epidemiol.* **Volume 11**, 753–767 (2019).

16. Weiskopf, N. G., Rusanov, A. & Weng, C. Sick Patients Have More Data: The Non-Random Completeness of Electronic Health Records. 6.

17. Nissen, F. *et al.* Validation of asthma recording in the Clinical Practice Research Datalink (CPRD). *BMJ Open* **7**, e017474 (2017).

18. Kroeker, K., Widdifield, J., Muthukumarana, S., Jiang, D. & Lix, L. M. Model-based methods for case definitions from administrative health data: application to rheumatoid arthritis. *BMJ Open* **7**, e016173 (2017).

19. Thygesen, K. *et al.* Fourth Universal Definition of Myocardial Infarction (2018). *Circulation* **138**, (2018).

20. Pathak, J., Kho, A. N. & Denny, J. C. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *J. Am. Med. Inform. Assoc.* **20**, e206–e211 (2013).

21. Ehrenstein, V., Nielsen, H., Pedersen, A. B., Johnsen, S. P. & Pedersen, L. Clinical epidemiology in the era of big data: new opportunities, familiar challenges. *Clin. Epidemiol.* **Volume 9**, 245–250 (2017).

22. Lix, L., De Coster, C., Currie, R. J., & Manitoba Centre for Health Policy. *Defining and validating chronic diseases: an administrative data approach*. (Manitoba Centre for Health Policy, 2006).

23. Gerber, Y., Weston, S. A., Jiang, R. & Roger, V. L. The changing epidemiology of myocardial infarction in Olmsted County, Minnesota, 1995-2012. *Am. J. Med.* **128**, 144–151 (2015).

24. Smolina, K., Wright, F. L. & Rayner, M. Determinants of the decline in mortality from acute myocardial infarction in England between 2002 and 2010 : linked national database study. *BMJ* **344**, 1–9 (2012).

25. Rapsomaniki, E. *et al.* Using big data from health records from four countries to evaluate chronic disease outcomes : a study in 114 364 survivors of myocardial infarction. *Eur. Heart J.* **2**, 172–183 (2016).

26. Payne, R. A., Abel, G. A. & Simpson, C. R. A retrospective cohort study assessing patient characteristics and the incidence of cardiovascular disease using linked routine primary and secondary care data. *BMJ Open* **2**, 1–8 (2012).

27. Nedkoff, L. *et al.* Identification of myocardial infarction type from electronic hospital data in England and Australia: A comparative data linkage study. *BMJ Open* **7**, 1–6 (2017).

28. Tran, J. *et al.* Patterns and temporal trends of comorbidity among adult patients with incident cardiovascular disease in the UK between 2000 and 2014: A population-based cohort study. *PLoS Med.* **15**, 1–23 (2018).

29. Conrad, N. *et al.* Temporal trends and patterns in heart failure incidence : a population-based study of 4 million individuals. *Lancet* **391**, 572–580 (2018).

TABLES

Table 2. Variables and Datasets for Applying MI Phenotyping Algorithms in the ARIC Study Data

	Item	Variable	Dataset	Description
ARIC Classification	Definite, Probable, Suspect, No MI	CMIDX	C18EVT1	Final MI classification by MMCC or computer algorithm if MMCC review not required
	ECG evidence	CECGDXX	C18OCC1	1 = absent, Uncodable, other; 2 = equivocal; 3 = evolving ST-T; 4 = diagnostic; 5 = evolving diagnostic
	Biomarker evidence	CENZDX2	C18OCC1	Downgraded; 1 = normal; 2 = incomplete; 3 = equivocal, 4 = abnormal
	Chest pain symptoms	CPAINDX2	C18OCC1	Downgraded; 1 = pain is absent or pain is present and of non-cardiac origin; 2 = pain of cardiac origin
Algorithm 2A	(I21 or I22) in any position in hospital discharge list	CELB10A through CELB10Z3	C18CELB1	All discharge diagnoses from hospitalization recorded
Algorithm 2B	(I21 or I22) in primary or secondary position in hospital discharge list	CELB10A, CELB10B	C18CELB1	Primary and secondary discharge codes
Algorithm 3A	(I21 or I22) in any position in hospital discharge list *AND*	CELB10A through CELB10Z3	C18CELB1	All discharge diagnoses from hospitalization recorded
	Elevated cardiac biomarker (troponin I, troponin T, CK-MB)	HRAA20E3	C18HRMA1	Cardiac enzymes above normal limit
		CENZDX2=4	C18OCC1	Downgraded to account for other reasons for elevated cardiac enzymes
	OR cardiac procedure during hospitalization	HRAA29C	C18HRMA1	Coronary angioplasty
		HRAA29C2	C18HRMA1	Coronary atherectomy
		HRAA29F	C18HRMA1	Coronary CT
		HRAA29P1	C18HRMA1	Coronary stent
Algorithm 3B	(I21 or I22) in primary or secondary position in hospital discharge list *AND*	CELB10A, CELB10B	C18CELB1	Primary and secondary discharge codes
	Elevated cardiac biomarker (troponin I, troponin T, CK-MB)	HRAA20E3	C18HRMA1	Cardiac enzymes above normal limit
		CENZDX24	C18OCC1	Downgraded to account for other reasons for elevated cardiac enzymes
	OR cardiac procedure during hospitalization	HRAA29C	C18HRMA1	Coronary angioplasty
		HRAA29C2	C18HRMA1	Coronary atherectomy
		HRAA29F	C18HRMA1	Coronary CT
		HRAA29P1	C18HRMA1	Coronary stent

Table 3. Variables and Datasets for Applying Heart Failure Phenotyping Algorithms in the ARIC Study Data

	Item	Variable	Dataset	Description
ARIC Classification	Definite ADHF (A), probable ADHF (B), chronic stable HF (C), unlikely HF (D), unclassifiable (E)	CHFDIAG	HFC18OCC1	MMCC adjudicated HF diagnosis
	Definite or probable AHDF (A or B), Chronic stable HF (C), Unlikely or unclassifiable (D or E)	CHFIDAG3	HFC18OCC1	Values 1 = A or B, 2 = C, 3 = D or E
Other HF Criteria	Framingham Criteria	FRAMINGHAM	HFC18OCC1	NPR (not present); PRS (HF present)
	Gothenburg Criteria	GOTHENBURG	HFC18OCC1	0 (absent) 1 (latent) 2 (manifest) 3 (grade 3) 4 (hf death) 5 (unknown)
	Modified Boston Criteria	MBOSTON	HFC18OCC1	DEF (definite), POS (Possible), UNLK (unlikely)
	NHANES Criteria	NHANES	HFC18OCC1	NPR (not present); PRS (HF present)
	Trialist Criteria	TRIALISTHF	HFC18OCC1	0, 1
Algorithm 2A	(I50, I13.0, I13.2, or I11.0) in any position in hospital discharge list	CELB10A through CELB10Z3	C18CELB1	All discharge diagnoses from hospitalization recorded
Algorithm 2B	(I50, I13.0, I13.2, or I11.0) in primary or secondary position in hospital discharge list	CELB10A, CELB10B	C18CELB1	Primary and secondary discharge codes
Algorithm 3A	(I50, I13.0, I13.2, or I11.0) in any position in hospital discharge list *AND*	CELB10A through CELB10Z3	C18CELB1	All discharge diagnoses from hospitalization recorded
	inpatient administration of IV diuretics	HFAA73B	C18HFAA1	
	OR (elevated BNP >500 pg/mL	HFAA39A	C18HFAA1	Worst BNP value
		HFAA39B	C18HFAA1	Last BNP value during hospitalization
		HFAA39C	C18HFAA1	BNP test upper limit normal (reference)
	or elevated NT-proBNP >450 pg/mL or >900 pg/mL for those <50 years* and ≥ 50 years, respectively)	HFAA40A	C18HFAA1	Worst NT-proBNP value
		HFAA40B	C18HFAA1	Last NT-proBNP value during hospitalization
		HFAA40C	C18HFAA1	NT-proBNP test upper limit normal (reference)
Algorithm 3B	(I50, I13.0, I13.2, or I11.0) in primary or secondary position in hospital discharge list	CELB10A, CELB10B	C18CELB1	Primary and secondary discharge codes
	inpatient administration of IV diuretics	HFAA73B	C18HFAA1	
		HFAA39A	C18HFAA1	Worst BNP value

	Item	Variable	Dataset	Description
	OR (elevated BNP >500 pg/mL	HFAA39B	C18HFAA1	Last BNP value during hospitalization
		HFAA39C	C18HFAA1	BNP test upper limit normal (reference)
	or elevated NT-proBNP >450 pg/mL or >900 pg/mL for those <50 years* and ≥ 50 years, respectively)	HFAA40A	C18HFAA1	Worst NT-proBNP value
		HFAA40B	C18HFAA1	Last NT-proBNP value during hospitalization
		HFAA40C	C18HFAA1	NT-proBNP test upper limit normal (reference)

ADHF: acute decompensated heart failure; HF: heart failure;

Table 4. Phenotypic Comparison Table for 4 MI Algorithms *within* each ARIC MI Classification Category

	Definite/Probable MI*				Suspect/No MI*			
	Algorithm 2A Numerator	Algorithm 2B Numerator	Algorithm 3A Numerator	Algorithm 3B Numerator	Algorithm 2A Numerator	Algorithm 2B Numerator	Algorithm 3A Numerator	Algorithm 3B Numerator
N (%)								
Age, years								
Age category[†]								
74 – 84 years								
85 years and over								
Women								
Race								
White								
Black								
Race-Gender								
White Men								
White Women								
Black Men								
Black Women								
Center								
Jackson, MS								
Forsyth Co., NC								
Minneapolis, MN								
Washington Co., MD								
Smoking status								
Current								
Former								
Never								
Unknown								
Missing								
Comorbidities								
Hypertension								
Diabetes								
Kidney disease								
Kidney failure								
Mortality								
Hospitalization								
30-day								

1-year								
Coexisting Cardiovascular Disease								
Heart failure								
Atrial fibrillation								
Stroke/TIA								
Severity Indicators								
STEMI								
NSTEMI								
Unclassified MI Type								
Cardiogenic Shock								
MI within 28 days of previous event								
Acute stroke during hospitalization								
Acute HF during hospitalization								

*Table will also be completed for each of the 4 MI classifications separately (definite MI, probable MI, suspect MI, unlikely MI);

[†]In 2016, the youngest possible age of an ARIC participant was 74 years of age

Table 5. Phenotypic Comparison Table for 4 Heart Failure Algorithms *within* each ARIC Heart Failure Classification Category

	Definite/Probable Acute Decompensated Heart Failure*				**Chronic Stable Heart Failure***			
	Algorithm 2A Numerator	Algorithm 2B Numerator	Algorithm 3A Numerator	Algorithm 3B Numerator	Algorithm 2A Numerator	Algorithm 2B Numerator	Algorithm 3A Numerator	Algorithm 3B Numerator
N (%)								
Age, years								
Age category[†]								
74 – 84 years								
85 years and over								
Women								
Race								
White								
Black								
Race-Gender								
White Men								
White Women								
Black Men								
Black Women								
Center								
Jackson, MS								
Forsyth Co., NC								
Minneapolis, MN								
Washington Co., MD								
BMI (kg/m^2) (mean, SD)								
BMI ≥ 30 kg/m^2								
Missing								
Smoking status								
Current								
Former								
Never								
Unknown								
Missing								
Comorbidities								
Hypertension								
Diabetes								
Kidney disease								
Kidney failure								
Dialysis								
Chronic bronchitis or COPD								
Asthma								
History of pulmonary embolism								
Mortality								

Hospitalization								
30-day								
1-year								
Coexisting Cardiovascular Disease								
Previous MI								
Ischemic Heart Disease								
Ischemic Cardiomyopathy								
Idiopathic or dilated cardiomyopathy								
Other cardiomyopathy								
Atrial fibrillation								
Stroke/TIA								
Severity Indicators								
Ejection Fraction (%)								
Ejection Fraction < 50%								
Ejection Fraction < 30%								
Previous CABG								
Previous PCI								
Previous Valvular Surgery								
Pacemaker								
Implantable Defibrillator								
HF diagnosis on record prior to index hospitalization								
Previous HF hospitalization prior to index hospitalization								
HF treatment documented prior to index hospitalization								
Acute on Chronic HF								

ICD-10-CM Codes								
Chronic HF ICD-10-CM Codes								

*Table will also be completed for Definite and Probable Acute Decompensated Heart Failure separately, and for Unlikely Heart Failure and Unclassifiable Heart Failure; [†]In 2016, the youngest possible age of an ARIC participant was 74 years of age. Acute on chronic HF ICd-10-CM codes include I50.23, I50.33, I50.43, and I50.813. Chronic HF ICD-10-CM codes include 150.22, 150.32, 150.42, and 150.812.

Table 6. Data Sources for MI Phenotyping Algorithm versus ARIC Classification Phenotypic Comparisons

Item	Variable	Dataset	Description
Age, years	CEVTDAT3	C18EVT1	Calculated: event date – date of birth
	DOB	C18CELB1	
Women	SEX	C18OCC1	Gender associated with hospitalization
	GENDER71	DERIVE71	Gender recorded at visit 1
Race	Race1	C18OCC1	Race associated with hospitalization
	RACEGRP71	DERIVE71	Race group recorded at visit 1
Center	CENTER	C18EVT1	
	CENTER	DERIVE71	Should match Center in C18EVT1
Smoking status	HRAA21D	C18HRMA1	Smoking status as recorded at hospitalization
	CURSMK72, FORSMK72, EVRSMK72, CIGT72	DERIVE71	Smoking status recorded at visit 7
BMI	BMI71	DERIVE71	From visit 7
Comorbidities			
Hypertension	HRAA38	C18HRMA1	History of hypertension recorded at hospitalization
	HYPERT75	DERIVE71	SBP ≥ 140 or DBP ≥ 90 or anti-hypertension medication recorded at visit 7
Diabetes	HRAA38B	C18HRMA1	Recorded at hospitalization
	DIABTS75	DERIVE71	Fasting blood glucose ≥ 126 mg/dL or non-fasting glucose ≥ 200 mg/dL or using medication for diabetes at visit 7
Kidney disease	Inc_ckd_defy_vx	INC_CKD_BY##	Incident CKD definition y between visit x and year ##
	EGFRCR71	DERIVE71	eGFR-Cr measured at visit 7 (<60)
Kidney failure	Inc_kf_vx	INC_KF_BY##	Incident kidney failure from visit X through year ##
	EGFRCR71	DERIVE71	eGFR-Cr measured at visit 7 (<15)
Mortality	DATED18 CEVTDAT3 C7_DATEMI	INCBY18 C18EVT1 INCBY18	Death date MI date MI date
Coexisting Cardiovascular Disease			
Heart failure	C7_INCHF_P_V5	INCBY18	Hospitalized HF with V5 as baseline
	C7_DATE_INCHF_P_V5	INCBY18	Date of first incident heart failure post visit 5
	C7_INCHF18	INCBY18	Incident HF (or death due to HF) by ICD code and no prevalent HF at visit 1
	C7_DATE_INCHF18	INCBY18	Date of first incident heart failure
	PREVHF01	INCBY18	Prevalent heart failure at visit 1
Incident heart failure following MI	C7_INCHF_P_MI	INCBY18	Missing if MI before incident HF
	C7_DATE_INCHF_P_MI	INCBY18	Date of first incident heart failure post MI
Atrial fibrillation	INCSELFREPAF INCSELFREPAF_DATE	STATUS7#	Where # = version number Self-report AF date or last FU date prior to end of visit 7
Stroke/TIA	HRAA39	C18HRMA1	History of stroke noted in medical record

Item	Variable	Dataset	Description
	TIAB01	INCBY18	History of stroke or TIA reported at visit 1
	C7_IN18ISC	INCBY18	Definite or probable incident ischemic stroke before CENSDAT7; use C7_ED18ISC (date of stroke admission) and EVTDAT (MI date)
	C7_ED18ISC	INCBY18	Hospital admission date for stroke or censoring date for non-incident events
	PRVSTR71	DERIVE71	Prevalent stroke by end of visit 7
Severity Indicators			
STEMI	CSTEMI	C18EVT1	
NSTEMI	CNSTEMI	C18EVT1	
Unclassified MI Type	MI3, CSTEMI, CNSTEMI	C18EVT1	MI3 = 1 and CSTEMI = 0 and CNSTEMI = 0
Cardiogenic Shock	HRAA28a	C18HRMA1	
MI within 28 days of previous event	C_LINK	C18OCC1	C_link = 1 if MI occurrence is linked with another MI occurrence within 28 days
Acute stroke during hospitalization	HRAA28G	C18HRMA1	
Acute HF during hospitalization	HRAA28B	C18HRMA1	

Table 7. Data Sources for HF Phenotyping Algorithm versus ARIC Classification Phenotypic Comparisons

Item	Variable	Dataset	Description
Age	HFEVTDATE	HFC18OCC1	Calculated: event date – date of birth
	DOB	C18CELB1	recorded at hospitalization
Gender	SEX	HFC18OCC1	Gender associated with hospitalization
	GENDER71	DERIVE71	Gender recorded at visit 1
Race	Race1	HFC18OCC1	Race associated with hospitalization
	RACEGRP71	DERIVE71	Race group recorded at visit 1
Center	CENTER	HFC18OCC1	recorded at hospitalization
	CENTER	DERIVE71	Should match Center in HFC18OCC1
BMI	BMI71	DERIVE71	From visit 7
	BMI	HFC18OCC1	BMI at discharge
Smoking status	CURSMK72, FORSMK72, EVRSMK72, CIGT72	DERIVE71	Smoking status recorded at visit 7
Comorbidities			
Hypertension	HFAA11J	C18HFAA1	History of hypertension recorded at hospitalization
	HYPERT75	DERIVE71	SBP ≥ 140 or DBP ≥ 90 or anti-hypertension medication recorded at visit 7
Diabetes	HFAA12A	C18HFAA1	Recorded at hospitalization
	DIABTS75	DERIVE71	Fasting blood glucose ≥ 126 mg/dL or non-fasting glucose ≥ 200 mg/dL or using medication for diabetes at visit 7
Kidney disease	Inc_ckd_defy_vx	INC_CKD_BY##	Incident CKD definition y between visit x and year ##
	EGFRCR71	DERIVE71	eGFR-Cr measured at visit 7 (<60)
Kidney failure	Inc_kf_vx	INC_KF_BY##	Incident kidney failure from visit X through year ##
	EGFRCR71	DERIVE71	eGFR-Cr measured at visit 7 (<15)
	HFAA13A	C18HFAA1	*Dialysis at hospitalization*
Chronic bronchitis or COPD	HFAA10B	C18HFAA1	recorded at hospitalization
Asthma	HFAA10A	C18HFAA1	recorded at hospitalization
History of pulmonary embolism	HFAA10D	C18HFAA1	recorded at hospitalization
Mortality	DATED18 HFEVTDATE C7_DATEINCHF18	INCBY18 HFC18OCC1 INCBY18	Death date HF date HF date
Coexisting Cardiovascular Disease			
Previous MI	HFAA11K	C18HFAA1	recorded at hospitalization
CHD ever	HFAA11H	C18HFAA1	recorded at hospitalization
Ischemic Cardiomyopathy	HFAA6A	C18HFAA1	recorded at hospitalization
Idiopathic or dilated cardiomyopathy	HFAA6B	C18HFAA1	recorded at hospitalization
Other cardiomyopathy	HFAA6I	C18HFAA1	recorded at hospitalization
Atrial fibrillation	INCSELFREPAF INCSELFREPAF_DATE	STATUS7#	Where # = version number

Item	**Variable**	**Dataset**	**Description**
			Self-report AF date or last FU date prior to end of visit 7
	HFAA11B1	C18HFAA1	Atrial fibrillation or flutter recorded at hospitalization
Stroke/TIA	HFAA14A	C18HFAA1	Recorded at hospitalization
	TIAB01	INCBY18	History of stroke or TIA reported at visit 1
	C7_IN18ISC	INCBY18	Definite or probable incident ischemic stroke before CENSDAT7; use C7_ED18ISC (date of stroke admission) and EVTDAT (MI date)
	C7_ED18ISC	INCBY18	Hospital admission date for stroke or censoring date for non-incident events
	PRVSTR71	DERIVE71	Prevalent stroke by end of visit 7
Severity Indicators			
Ejection Fraction (%)	LVEF_CUR	HFC18OCC1	*Current EF*
Ejection Fraction < 50%	LVEF_CUR_LOW	HFC18OCC1	*Current EF categorized as < 50 or ≥50*
Ejection Fraction < 30%	*Calculated*	HFC18OCC1	recorded at hospitalization
Previous CABG	HFAA11E1	C18HFAA1	recorded at hospitalization
Previous PCI	HFAA11E2	C18HFAA1	recorded at hospitalization
Previous Valvular Surgery	HFAA11E3	C18HFAA1	recorded at hospitalization
Pacemaker	HFAA11E4	C18HFAA1	recorded at hospitalization
Implantable Defibrillator	HFAA11E5	C18HFAA1	recorded at hospitalization
HF diagnosis on record prior to index hospitalization	HFAA7A	C18HFAA1	recorded at hospitalization
Previous HF hospitalization prior to index hospitalization	HFAA7B	C18HFAA1	recorded at hospitalization
HF treatment documented prior to index hospitalization	HFAA7C	C18HFAA1	recorded at hospitalization

Table 8. Distribution of ARIC MI Data for Determining MI Diagnosis by MI Phenotyping Algorithm

	Algorithm 2A	Algorithm 2B	Algorithm 3A	Algorithm 3B
Biomarker Evidence				
Abnormal				
Equivocal				
Incomplete				
Normal				
ECG Evidence				
Evolving diagnostic				
Evolving ST-T				
Equivocal				
Absent or Uncodable				
Chest pain of cardiac origin				
Present				
Absent				

Biomarkers include troponin I, troponin T, and CK-MB; ST-T refers to ST-segment and T-waves in ECG; ECG: electrocardiogram; chest pain of cardiac origin determined from downgraded chest pain symptom classification

Table 9. Describing Heart Failure Hospitalization by Phenotyping Algorithm and ARIC Heart Failure Classification

	Definite/Probable Acute Decompensated Heart Failure				**Chronic Stable Heart Failure**			
	Algorithm 2A	Algorithm 2B	Algorithm 3A	Algorithm 3B	Algorithm 2A	Algorithm 2B	Algorithm 3A	Algorithm 3B
Transthoracic echocardiogram performed during hospitalization								
Left ventricular hypertrophy								
Pulmonary hypertension								
Dilated left ventricle								
Dilated right ventricle								
Diastolic dysfunction								
Impaired left ventricle systolic function								
Impaired right ventricle systolic function								
Aortic regurgitation								
Aortic stenosis								
Tricuspid regurgitation								
Mitral regurgitation								
Mitral stenosis								
Transesophageal echocardiogram performed during hospitalization								
Dilated left ventricle								
Dilated right ventricle								
Impaired left ventricle systolic function								

Impaired right ventricle systolic function								
Coronary angiography performed								
Previous CABG grafts present								
Number of occluded grafts*								
0								
1								
2								
3								

Table 10. Codes for Incident Chronic Kidney Disease Stage 3+ (Definition 2), The ARIC Study

ICD-9-code	Description	ICD-10-code
582	Chronic glomerulonephritis	N03
583	Nephritis and nephropathy	
585, 585.x where x≥3	Chronic kidney disease	N18, N18.x where x≥3
586	Kidney failure	N19
587	Kidney sclerosis	N26
588	Disorders resulting from impaired Kidney function	N25
403	Hypertensive chronic kidney disease	I12
404	Hypertensive heart and kidney disease	I13
593.9	Unspecified disorder of the kidney and ureter	
250.4	Diabetes with Kidney complications	E10.2, E11.2, E13.2
V42.0	Kidney replaced by transplant	Z94.0
55.6	Transplant of kidney	
996.81	Complications of transplanted kidney	
V45.1	Kidney dialysis status	Z99.2
V56	Admission for dialysis treatment or session	Z49
39.95	Hemodialysis	
54.98	Peritoneal dialysis	
	Encounter for adjustment and management of vascular access device	Z45.2

*Codes in gray rows counted as incident kidney disease only if a concomitant acute kidney injury code (ICD-9: 584.x, ICD-10-: N17) is not present

Source: Derived and Incident Kidney Disease Documentation (Section V), The ARIC Study, Updated January 18 2019.

Table 11. Codes for Incident Kidney Failure, The ARIC Study

ICD-9-code	Description	ICD-10-code
V42.0	Kidney replaced by transplant	Z94.0
55.6	Transplant of kidney	
996.81	Complications of transplanted kidney	
V45.1	Kidney dialysis status	Z99.2
V56	Admission for dialysis treatment or session	Z49

39.95	Hemodialysis	
54.98	Peritoneal dialysis	
	Encounter for adjustment and management of vascular access device	Z45.2
585.5	Chronic kidney disease stage 5	N18.5
585.6	End stage Kidney disease	N18.6
586	Kidney failure	N19
403.01	Hypertensive chronic kidney disease, malignant, with CKD 5 or ESRD	
403.91	Hypertensive chronic kidney disease, with CKD 5 or ESRD	I12.0

*Codes in gray rows not counted as incident kidney failure if for hospitalizations a concurrent AKI code (ICD-9: 584.x, ICD-10-: N17) is present or for deaths, a concurrent AKI code is present without a concurrent CKD code.

Source: Derived and Incident Kidney Disease Documentation (Section VI), The ARIC Study, Updated January 18 2019.

Table 12. Contingency Table for MI Algorithms and Binary MI ARIC Classification

		ARIC Cohort Surveillance Classification	
		Definite/Probable MI	Suspect MI/No MI
MI Algorithm 2A	MI		
	No MI		
MI Algorithm 2B	MI		
	No MI		
MI Algorithm 3A	MI		
	No MI		
MI Algorithm 3B	MI		
	No MI		

Table 13. Contingency Table for HF Algorithms and Binary HF ARIC Classification

		ARIC Cohort Surveillance Classification	
		A, B or C	D or E
HF Algorithm 2A	HF		
	No HF		
HF Algorithm 2B	HF		
	No HF		
HF Algorithm 3A	HF		
	No HF		
HF Algorithm 3B	HF		
	No HF		
A = definite acute decompensated HF, B = probable acute decompensated HF, C = chronic stable HF, D = unlikely HF, E = unclassifiable			

Table 14. Contingency Table for MI Algorithms and Four MI ARIC Classifications

		ARIC Cohort Surveillance Classification			
		Definite MI	Probable MI	Suspect MI	No MI
MI Algorithm 2A	MI				
	No MI				
MI Algorithm 2B	MI				
	No MI				
MI Algorithm 3A	MI				
	No MI				
MI Algorithm 3B	MI				
	No MI				

Table 15. Contingency Table for HF Algorithms and Five HF ARIC Classifications

		ARIC Cohort Surveillance Classification				
		A	B	C	D	E

HF Algorithm 2A	HF					
	No HF					
HF Algorithm 2B	HF					
	No HF					
HF Algorithm 3A	HF					
	No HF					
HF Algorithm 3B	HF					
	No HF					

A = definite acute decompensated HF, B = probable acute decompensated HF, C = chronic stable HF, D = unlikely HF, E = unclassifiable

Table 16. Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value for Acute Myocardial Infarction Algorithms compared to ARIC Hospitalized Myocardial Infarction Classifications

	Algorithm 2A	Algorithm 2B	Algorithm 3A	Algorithm 3B
Definite MI				
Sensitivity	x.x (95% CI: x.x, x.x.)			
Specificity				
PPV				
NPV				
Probable MI				
Sensitivity				
Specificity				
PPV				
NPV				
Suspect M				
Sensitivity				
Specificity				
PPV				
NPV				
No MI				
Sensitivity				
Specificity				
PPV				
NPV				
Definite/Probable MI				
Sensitivity				
Specificity				
PPV				
NPV				
Suspect/No MI				
Sensitivity				
Specificity				
PPV				
NPV				

MI = myocardial infarction; PPV = positive predictive value; NPV = negative predictive value

Table 17. Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value for Heart Failure Algorithms compared to ARIC Hospitalized Heart Failure Classifications

	Algorithm 2A	Algorithm 2B	Algorithm 3A	Algorithm 3B
Definite Acute Decompensated Heart Failure				
Sensitivity	x.x (95% CI: x.x, x.x)			
Specificity				
PPV				
NPV				
Probable Acute Decompensated Heart Failure				
Sensitivity				
Specificity				
PPV				
NPV				
Chronic Stable Heart Failure				
Sensitivity				
Specificity				
PPV				
NPV				
Unlikely Heart Failure				
Sensitivity				
Specificity				
PPV				
NPV				
Unclassifiable				
Sensitivity				
Specificity				
PPV				
NPV				
Definite/Probable Acute Decompensated Heart Failure				
Sensitivity				
Specificity				
PPV				
NPV				
Unlikely/Unclassifiable Heart Failure				
Sensitivity				
Specificity				
PPV				
NPV				

PPV = positive predictive value; NPV = negative predictive value

Table 18. Formulas for Sensitivity, Specificity, PPV, and NPV with Binary MI Classification

		ARIC Cohort Surveillance Classification		**Algorithm SN, SP, PPV, and NPV formulas**	
		Definite / Probable MI	Suspect MI / No MI		
MI Algorithm 2A	MI	a_1	b_1	*SN*	$\frac{a_1}{a_1 + c_1}$
				SP	$\frac{d_1}{b_1 + d_1}$
	No MI	c_1	d_1	*PPV*	$\frac{a_1}{a_1 + b_1}$
				NPV	$\frac{d_1}{c_1 + d_1}$
MI Algorithm 2B	MI	a_2	b_2	*SN*	$\frac{a_2}{a_2 + c_2}$
				SP	$\frac{d_1}{b_2 + d_2}$
	No MI	c_2	d_2	*PPV*	$\frac{a_2}{a_2 + b_2}$
				NPV	$\frac{d_2}{c_2 + d_2}$
MI Algorithm 3A	MI	a_3	b_3	*SN*	$\frac{a_3}{a_3 + c_3}$
				SP	$\frac{d_3}{b_3 + d_3}$
	No MI	c_3	d_3	*PPV*	$\frac{a_3}{a_3 + b_3}$
				NPV	$\frac{d_3}{c_3 + d_3}$
MI Algorithm 3B	MI	a_4	b_4	*SN*	$\frac{a_4}{a_4 + c_4}$
				SP	$\frac{d_4}{b_4 + d_4}$
	No MI	c_4	d_4	*PPV*	$\frac{a_4}{a_4 + b_4}$
				NPV	$\frac{d_4}{c_4 + d_4}$

MI: myocardial infarction; SN: sensitivity; SP: specificity; PPV: positive predictive value; NPV: negative predictive value.