Manual 3.b

EHR-based Surveillance of Hospitalized Myocardial Infarction and Heart Failure

ARIC EHR Feasibility Study

Version 1.0
April 28, 2017

Glossary

CDART: Carolina Data Acquisition and Reporting Tool (research data management system)

CDM: Common Data Model

CHD: Coronary Heart Disease

cTAKES™: Apache cTAKES™ - clinical Text Analysis Knowledge Extraction System

CUI: Unified Medical Language System Concept Unique Identifier

CDW: Clinical Data Warehouse

EHR: Electronic Health Record(s)

HF: Heart Failure

H-list: List of hospital records sampled for abstraction by ARIC personnel

HRN: Hospital Record Number

MI: Myocardial Infarct

NLP: Natural Language Processing

I. Overview

The Atherosclerosis Risk in Communities (ARIC) Study has conducted epidemiologic surveillance of four geographically defined areas from 1987 through 2015, to describe trends in cardiovascular disease. Eligible discharge diagnoses retrieved yearly from all hospitals in the four geographic locales were sampled by ARIC and hospital records abstracted by ARIC study personnel. These items were then used to classify hospitalized myocardial infarction and heart failure according to ARIC's study criteria, as detailed in the corresponding manuals of procedures (https://www2.cscc.unc.edu/aric/surveillance-manuals).

As of November 2016 ARIC is conducting a pilot study of automated information extraction from electronic hospital health records (EHR), using software developed and calibrated by the ARIC study investigators. Information extracted from EHR using computer algorithms is then compared to data abstracted from the same hospital records by ARIC's trained personnel following the standardized protocol.

The automated information extraction from EHR targets the same data items, and as much as is possible, applies the same extraction rules as those used by ARIC abstractors. Data elements extracted from the EHR include information recorded as text and also data elements in pre-coded formats (structured data elements). In contrast to the standard record abstraction by ARIC, the automated information extraction is done centrally at the ARIC Coordinating Center (CC). For this, hospitals that provide access to their medical records by the ARIC personnel are asked to retrieve sections of the EHR from their clinical data repositories/data warehouses and provide a copy to the CC.

To obtain complete and accurate copies of the portions of the medical record that ARIC abstractors normally access, clear guidelines for the retrieval of data elements from EHR are provided to each clinical data repository/clinical data warehouse (CDW). The process by which EHR are requested of a CDW, the transfer, management and de-identification of EHR, as well as the extraction of data elements from EHR and the assessment of agreement with data abstracted visually by ARIC are described in this manual of procedures.

II. Study Design Features

Since its inception in 1989 ARIC surveillance monitored and validated hospitalized myocardial infarcts among ARIC cohort members and among age-eligible residents of four geographically define study areas: Forsyth County, North Carolina; Jackson, Mississippi; selected suburbs of Minneapolis, Minnesota; and Washington County, Maryland. Since January 1, 2005, the ARIC study enumerated and validated hospitalized heart failure among members of the ARIC cohort and among men and women aged 55 years and above, among residents of the four ARIC study communities. The standardized study procedures are described in ARIC manuals 3 and 3.a, respectively, located at https://www2.cscc.unc.edu/aric/surveillance-manuals

Each year hospitals in the ARIC Surveillance network were asked for discharge index listings of records according to selection criteria provided by ARIC. All hospitalizations of ARIC cohort members and a sample of hospitalization records of the surveillance communities drawn by the ARIC coordinating center (CC) for each hospital were then abstracted by ARIC personnel,

adhering to standardized information extraction rules. Following November 15, 2016, the ongoing retrieval and abstraction of hospital records by the ARIC study is limited to the ARIC cohort. In parallel, this study is conducted over the course of three years to assess the feasibility of community surveillance based on EHR.

III. Selection of the Hospitals and Human Abstraction

1. Hospitals Selected

Of 20 hospitals participating in the ARIC cohort and community surveillance network in 2016, vanguard hospitals at each ARIC study site were selected for full participation in this feasibility study by providing EHR of the hospital records sampled and abstracted by ARIC. Vanguard hospitals were selected to achieve diversity in the EHR platforms accessed by this study. Of the major EHR platforms currently in use the following are accessed through the hospitals selected: EpicCare, McKesson and Allscripts. The feasibility study requests EHR for all 2015 ARIC cohort events (already abstracted), and for a sample of 2015 community hospitalizations at six of the hospitals in the ARIC study areas. The vanguard hospitals at each study site are identified in Table 1. These hospitals are: University of Mississippi Medical Center, Mississippi Baptist Medical Center, North Memorial Hospital, University of Minnesota Medical Center, Meritus Medical Center, and North Carolina Baptist Medical Center.

The numeric target is to abstract 212 CHD-eligible (community) hospitalizations and 125 heart failure-eligible (community) hospitalizations at each ARIC study site. The ARIC CC selects these hospitalizations from the discharge listings provided by each hospital. These hospitalizations are provided with event numbers and made available to ARIC personnel in CDART following a central abstractor training by webinar.

The number of hospital records targeted by this feasibility study includes 850 hospitalized CHD and 500 hospitalized heart failure community surveillance events sampled from the six vanguard hospitals, stratified by study site and hospital (EHR platform). These hospital records are abstracted by ARIC personnel following the standard ARIC protocol, and the EHR pertaining to these events are requested from the respective hospitals. EHR are also requested for all cohort events discharged from these vanguard hospitals during 2015, i.e., approximately 551 hospitalized HF and 1082 CHD cohort events already abstracted by ARIC at these vanguard hospitals.

The procedures by which hospital records are sampled are detailed below. Vanguard hospitals are contacted on an expedited time line for participation in this feasibility study, starting in December 2016. The remaining hospitals in the ARIC surveillance network are approached in the course of this feasibility study to assess their capabilities and willingness to provide copies of EHR to the ARIC study at a later stage.

Table 1. ARIC EHR surveillance feasibility study. Vanguard hospitals and sample of 2015 community surveillance records to be abstracted

#	Hospital	EHR System	Av. CHD Abst./year	Comm. to Abstract	Av. HF Abst/year	Comm. To Abstract	Retrieve 2015 EHR for:	
							Comm.	Cohort
a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*
11	NC Baptist	Epic	825	212	535	125	Yes	Yes
12	Forsyth MC	Epic	1065		660		No	No
14	Kernersville	Epic	~40		~30		No	No
15	Clemmons	Epic	50		55		No	No
21	U Miss MC	Epic	200	106	115	40	Yes	Yes
22	VA MC	Vista	60		60		No	No
23	St. Dominic's	Kronos	425		360			
24	Merit Health	McKesson	210		140			
25	Miss. Baptist MC	McKesson	380	106	305	85	Yes	Yes
26	River Oaks	McKesson	9		7		No	No
	Alina (30,35,44)	Epic + Other	20-90		7-40		No	No
32	Fairview Southd.	Epic + Fairv.	23		7		No	No
34	Hennepin Cty	Epic	40		18		No	No
36	Park Nicollet	Epic	170		110		No	No
40	North Memorial	Epic	565	200	307	120	Y	Y
45	U Minn MC	Epic + Fairv.	35	12	13	5	Y	Y
51	Meritus MC	Allscripts	1090	212	606	125	Y	Y
53	VA MC	--	15		14		No	No
54	U. Maryland	Cerner/Epic	40		13		No	No
57	Washington Hosp	MRDI	18		7		No	No

2. Recruitment of the Hospitals

Increasingly, hospitals have developed EHR capabilities managed through clinical data warehouses (CDW) that are set up to varying degrees to perform data queries and record extraction. Hospitals are encouraged by the Centers for Medicare and Medicaid Services (CMS) to grow these capabilities toward optimization of medical care and data sharing. There is considerable heterogeneity in the progress made by different hospitals toward this goal, common standards for data structure and sharing are still under development, and a number of differences also exist between the EHR platforms adopted by different hospitals. An awareness of each hospital's capabilities and EHR output data structure is important to a successful engagement of the hospital in this study.

Hospitals or their CDW are approached by the local ARIC investigators, typically following the channels and contacts in place through their established collaboration with the ARIC Study. As part of this initial contact, an overview of the feasibility study and its aims are presented, and basic information is requested about the CDW's data extraction capabilities and procedures. A sample letter used for this purpose is found in Appendix I. Questions from the hospital/the CDW are addressed, and applications for IRB approval are submitted at this point if requested by the hospital or deemed appropriate by the local ARIC principal investigator. Following these steps the detailed request for EHR is submitted, with specification of the records and data elements to be retrieved, as described in Section IV of this protocol, and as shown in Appendix II.

3. Selection of the Hospitalized Events for Human Abstraction and EHR Extraction
All 2015 hospitalized cohort events were abstracted if eligible, following standard ARIC protocol. EHR of these cohort events are requested from the vanguard hospitals (see Table 1). Lists for each hospital identifying these records by hospital record number (HRN) and discharge date are prepared by the ARIC CC.

EHR are requested also for a sample of eligible 2015 community surveillance events at these 6 hospitals (Table 1, columns e, i). Note: The number of hospital records shown in this table are hospital records abstracted (not sampled) and exclude cohort abstractions. The ARIC CC draws the 2015 Community Supplement as a sample of eligible community events discharged from the vanguard hospitals, to abstract ~500 hospitalized heart failure and ~850 CHD community surveillance events, stratified by study site and hospital to achieve comparable numbers of CHD and of heart failure abstractions at each study site. The sampling accounts for the proportions of records not eligible or not located observed at each study site.

4. Study forms Abstracted for the ARIC EHR Feasibility Study
Only 2015 Community Surveillance hospitalizations are abstracted, as presented on the H-List. The standard ARIC study forms to be abstracted are: HRA, HFA, CHI, and CFD. A NOF is completed if the hospitalization is not eligible. The supplemental HFA form is not used, nor is the DTH form used for the EHR feasibility study

5. Abstraction Protocol
The ARIC HRA, HFA, CHI, CFD and NOF forms are completed following the standard ARIC protocol, with the exception of ECG retrievals, as described below.
Hospital transfer records are not linked as part of the EHR feasibility study. Accordingly, items 8.a-e of the CHI form need not be filled.

6. Case materials to be Retrieved and Transferred
Case materials for the abstracted 2015 Community Surveillance hospitalizations are retrieved and sent following the regular ARIC protocol and using established procedures. All identifiers in the case material PDF documents are blinded per standard ARIC protocol.

IV. Requesting EHR from the CDW
ARIC field centers request EHR using an individualized version of the centrally prepared set of materials described below. While most CDWs have procedures and forms in place to request EHR, these tend to be general in nature and lack the specificity of detail needed to extract the data elements required by the ARIC abstraction protocol. This specificity is provided in the set of materials submitted by ARIC to the CDW, which includes the following: (a) cover letter to CDW analysts outlining the request; (b) a spreadsheet that specifies each data element requested, its location and format, and formatted data output records as example for the analyst; (c) a questionnaire for the analyst to return as a part of output records.

1. Calibrating the CDM tool to the EHR platform
Prior to submitting the request for EHR to the CDW a copy of the Excel workbook was provided to the ARIC field center personnel familiar with the hospital's EHR to review and complete the "Abstractor" tab for each hospital included in the EHR feasibility study.

For each data element, “locations” in the clinical record were identified. These refer to text notes or section of the medical record, and were listed in the order that an ARIC abstractor may search for that data element according to the ARIC protocol. Using this as a guide the ARIC abstractors listed the location of each item in the clinical record of this specific hospital, and the order of locations followed at this hospital to search for this element, if different from the sequence specified in the protocol.

In a different column the ARIC abstractor listed whether the data item is a structured data element in the electronic record at this hospital. We define structured data element as one that is pre-formatted and does not need to be read and interpreted from text notes in the medical record. For example, dates, age, discharge codes, time stamps and lab assay results are structured data elements. Procedures may be considered structured data elements because they correspond to CPT codes in an electronic data warehouse.

Input from the ARIC abstractors informed the design and format of the ARIC CDM tool to request EHR, as presented to the CDW analysts. Abstractor input on the location of data elements in the clinical records of each hospital and the sequence followed by abstractors in retrieving them is also considered in programming the post-processing retrieval of these data elements from different hospitals’ EHR.

2. Use of the CDM tool by the CDW analyst

The request for EHR is structured as an Excel workbook with eight sheets outlining the requested data elements and text notes. (See Appendix II). The template of the letter to analysts is included in this manual as Appendix II. Zipped files that include the spreadsheet of data elements requested, the sample output files and the questionnaire for analyst feedback are available by request to arichelp@unc.edu, and correspond to the set of materials submitted to the CDW analyst: EHR_DataElements_ANALYST_04.18.2017.xlsx; and ARIC EHR Feasibility Study_Analyst Questionnaire_2017.04.18.docx. A separate folder inside the zip file, called “Sample ARIC Analyst Files”, contains: exampleOutput_excelVersion.xlsm, an Excel version of synthetic data provided as a visual aid for analysts in providing the data output, and ***.txt files (e.g., CONDITIONS.txt, DIAGNOSES.txt), with synthetic data in the format and output style (pipe delimited text files) for each file. A list of the records to be retrieved is provided to each hospital as a zipped, password-protected file that includes medical record number and date of discharge.

3. Input from the CDW analyst and preparation test records

With the request for EHR the CDW analyst is asked to review the contents of the Excel workbook and the data elements listed within each tab. The analyst is encouraged to forward any questions or request for clarification to the contact at the ARIC field center. Otherwise, preparation of a test record based on the first 10 encounters listed in the file provided by ARIC is requested.

4. Review of test data and feedback

CDW analysts are invited to process the initial 10 records of the list of MRN provided by ARIC (which are sorted to include equal numbers of CHD and heart failure hospitalizations) using the transfer mechanism mentioned below. Feedback on the completeness and accuracy of the test records is provided to the CDW analyst prior to processing all records requested.

5. Transfer of EHR to the ARIC coordinating center

The study provides a secure portal for the CDW analyst to upload the requested output to the ARIC Coordinating Center (CC) at the University of North Carolina at Chapel Hill. If preferred, the CDW analyst can provide the output records to the designated ARIC staff person on site for a secure transfer to the ARIC CC following the ARIC protocol.

V. Record De-identification

On arrival at the ARIC CC the records are assigned the corresponding HOSPID provided by the ARIC CC, multiple lines per note are combined into one text source, a NOTE_TYPE_ID, NOTE_AUTHOR_TYPE_ID, and NOTE_DEPT_ID is assigned, and an ARIC ID is assigned according to the medical record-ARIC ID crosswalk table provided by the ARIC Coordinating Center. Medical record columns are then removed, and identifiers are scrubbed with MIT DE-ID software supplemented with US Census/Birth Records to remove all names. A sample of records is extracted periodically for manual verification of de-identification

VI. Extraction of EHR Data Elements at the ARIC Coordinating Center

1. Text Mining.

An overview of the protocol to extract data elements from text is shown in Appendix III. Briefly, special characters that are not compatible with cTAKES are removed and the length of the number of characters is counted in each note and stored as a variable. An initial list of CTAKES concepts and negation is created as an editable table (csv file or equivalent) that can be read into cTAKES script. The process is repeated until benchmarks are satisfied. For each Unified Medical Language System Concept Unique Identifier (CUI), a new variable is created and occurrences are counted for each note. For each negation term/CUI combination, a new variable is created and a count of negation occurrences for each note is set up. For each variable in the above steps, a binary variable is created identifying CUI occurrences for each note.

Sensitivity (recall) and specificity (precision) are then calculated by comparison to the corresponding data element abstracted by ARIC staff from the same record. If sensitivity $\geq$95% and specificity $\geq$75% thresholds are met, the negation terms and CUIs are added to the master list; otherwise, the sources of discordance between NLP and the ARIC abstractors is investigated by accessing the medical records, and the above process is repeated.

2. Structured Data Elements

The data elements likely to be in a structured format (not free text) are listed in the ARIC CDM tool and specified in the Excel workbook (ARIC_EHR_DataElements.xlsx) submitted to the CDW. These data elements are listed in the first seven worksheets. Table 2 is an overview of the eight worksheets and their descriptions. The column labeled *"Patient Encounter-to-data rows"* in this table indicates whether receiving a single row of data per patient encounter (*one-to-one*) is anticipated, or if multiple rows of data for a single patient encounter (*one-to-many*) are anticipated instead.

Table 2. Data elements listed in the Excel workbook (ARIC_EHR_DataElements.xlsx)

Worksheet Title	Type of Data	Patient Encounter-to-data rows	Description
DEMOGRAPHICS_ ENCOUNTER	Structured	One-to-one	Items from various relational tables within the CDW (domains) that can be combined into a single row for a given patient/encounter
CONDITIONS	Structured	One-to-many	Conditions reported by the patient or medical history/conditions that are carried to the current encounter.
DIAGNOSIS	Structured	One-to-many	Diagnoses during the encounter
PROCEDURES	Structured	One-to-many	Procedures recorded during the encounter
LABS	Structured	One-to-many	Laboratory measures
VITALS	Structured	One-to-many	Vital measurements, such as blood pressure, weight
MEDICATIONS	Structured	One-to-many	Medications
NOTES	Unstructured	One-to-many	List of text notes that are requested in addition to the structured data elements

By specifying a comprehensive extraction of the above structured data elements from EHR, including their date- and time-stamps, the clinical measurements, laboratory assays and medications specified by the ARIC protocol for individual patients and encounters are selected during post-processing of the EHR at the ARIC CC. Individual signs, symptoms, ECGs, and laboratory assays are selected at that point with reference to the time of onset of the event or the time since hospital admission, as specified by the ARIC protocol. Similarly, the temporal evolution of clinical manifestations or of acute-phase analytes such as high sensitivity troponins or NT-proBNP is determined during post-processing according to the timing specifications in the protocol followed by ARIC abstractors.

VII. Extraction of Electrocardiographic Patterns from EHR

In addition to pain and cardiac biomarkers, ARIC's classification of a hospitalized myocardial infarction requires visually coded information from up to three electrocardiogram (ECG) images retrieved per eligible hospitalization (see https://www2.cscc.unc.edu/aric/surveillance-manuals). Per protocol, coding the ECG tracings according to the Minnesota Code is performed centrally, at the Epidemiological Cardiology Research Center (EPICARE), at Wake Forest University.

1. Extraction of Clinical Interpretation of ECG Patterns

To obviate the need for visual coding, the EHR feasibility study extracts the algorhythmically derived clinical interpretation generated by the ECG machines and stored in the EHR for use in event classification of hospitalized myocardial infarcts. Up to three ECGs (selected according to their timing since admission) are selected for each eligible hospitalized event, and the automated report generated by the ECG machine is extracted from EHR as a combination of structured data elements and text mining. ECGs read at EPICARE according to the established ARIC protocol are grouped by the ARIC CC into five Minnesota coding-based diagnostic categories for event classification; the algorhythmically derived clinical interpretation retrieved from the EHR will thus be grouped into one of the five criteria used by ARIC in the classification of a myocardial infarct.

2. ECG Calibration

The ECG calibration component of the EHR feasibility study compares the ECGs processed at EPICARE according to the standard ARIC protocol to the coding based on the clinical interpretation generated by ECG machines, as extracted from EHR. The dictionary of terms used by the main ECG manufacturers in the clinical interpretation/narrative is comprehensively annotated for concepts applicable to myocardial ischemia or infarction, and mapped to the ECG diagnostic groupings used for ARIC's classification ("evolving diagnostic, diagnostic ECG, evolving ST-T, equivocal, and absent/ uncodable/ other"). For the development phase of this study, ECG narratives are extracted by ARIC abstractors from the EHR of the 2015 hospitalized, CHD-eligible cohort events abstracted by ARIC, for which ECGs were processed at EPICARE according to ARIC's standard criteria. In parallel, a mapping of the clinical descriptors in the report generated by the ECG machines to the Minnesota Code-based classification groupings is developed and optimized under the leadership of Dr. Elsayed Soliman, Director of EPICARE. The validation phase of the ECG calibration will draw on the clinical interpretation generated by the ECG machines extracted from EHR by the software developed by this study, as outlined in section VI. Based on the performance of this knowledge-based approach, machine learning (ML) may be considered for the classification of information from the ECG.

3. Processing of ECGs for 2015 Community Surveillance hospitalizations

Contrary to the standard ARIC abstraction protocol, 12-lead ECGs are <u>not</u> retrieved for the 2015 Community Surveillance hospitalizations abstracted for the EHR feasibility study.

4. Processing of ECG narratives for the ARIC EHR feasibility study

As part of the 2015 cohort hospitalizations previously abstracted at the six hospitals, ECGs were retrieved and sent to the ECG Reading Center. For this EHR feasibility study the 2015 cohort hospitalizations at these six hospitals are accessed once more, and the printed ECG narrative corresponding to each of the ECG tracings sent earlier for central processing is now extracted as an individual PDF file. The ECG images are not extracted again as part of this study; only their printed narratives are extracted. Thus, the first, third and last ECGs are identified again for the abstracted 2015 cohort events, and their summary statistics and narrative interpretations are retrieved, provided with the Event ID and labeled first, third and last as described below.

Some of the cohort ECG tracings previously sent to EPICARE by the abstractors may have come from linked transfer records. Since records are not linked for this feasibility study, ECGs and narratives that are discrepant by date and time will be subsequently excluded from analysis.

5. Uploading of the ECG narratives to the ECG Reading Center

The ECG narratives are processed and uploaded to the ECG Reading Center following the procedures set out in Appendix IV of ARIC Manual 3 (ver 6.6), "Instructions for Sending ECGs to the ECG Reading Center." To differentiate the previously sent ECGs from the ECG narratives being sent now, the file naming convention is changed by using EHRN (for EHR Narrative) in place of COH in the file name. If cohort event 3456789 had a first and last ECG, the name given to the merged PDF file previously sent to the Reading Center was 'COH3456789FL'. The file name for the (combined) ECG narratives now being sent to the Reading Center would be 'EHRN3456789FL.' With the sole exception of this change in the PDF file name, all procedures

and conventions set out in steps 1 – 3 of Appendix IV of MOP 3 are followed for the ECG narratives.

6. Agreement with Visually Coded ECG

Agreement by ECG code grouping and predictive value will be estimated for algorithmically derived items retrieved from EHR versus the Minnesota Code-based ARIC diagnostic categories. Bias in event classification attributable to ECG data missing/misclassified by electronic vs. human abstraction will be quantified.

VIII. Agreement of Visual Abstraction and Automated Data Extraction

The combination of the ARIC forms included in the EHR feasibility study yields a total of 325 data items (ARIC variables) available for item-specific comparisons between human abstraction and extraction from EHR as structured data elements or through text mining. Approximately 1/3rd of these data elements are critical to event eligibility determination and/or for classification of CHD or heart failure events according to ARIC protocol. The latter set of data elements will be given priority for the optimization of item-specific completeness and accuracy during the study's development phase, but all items (variables) may be investigated for completeness and accuracy of data extraction from EHR during the validation phase.

1. Evaluation of Structured Data Elements

Structured data elements extracted from EHR will be compared to those abstracted by ARIC staff in terms of completeness, agreement, and predictive value.

2. Evaluation of Text Mining

Data elements extracted from EHR by means of NLP will be compared at the item (variable) level to those abstracted by ARIC staff in terms of completeness, sensitivity and specificity.

IX. Event Classification

ARIC's classification of a myocardial infarct or heart failure event is not modified as part of this EHR feasibility study. Instead, once standards of completeness and accuracy of data extraction from EHR have been met, variables originating from EHR extraction that are used to determine event eligibility or classification may serve as input to the standard procedures for event classification that are in place at the ARIC CC. The performance characteristics of the procedures to extract the desired information from EHR will be gauged at the data element level, whereas the event classification procedures are unchanged.

X. Bias in Measures of Event Occurrence

The ARIC EHR feasibility study is designed to develop, optimize and validate generalizable tools for extraction of data elements from the contemporaneous EHR supported by different provider platforms. Its success in meeting these aims is measured by the performance of the data extraction tools developed and their cost-efficiency, and their potential impact on the wider use of EHR in quantifying the magnitude of health events in populations and their temporal trends. The characteristics of the error associated with data element extraction from EHR vs. human abstractors is quantified in this study and can inform the estimation of bias in event rates associated with the use of EHR according to the protocol developed by thus study. This EHR feasibility study is not powered to provide precise estimates of such bias, however.

Appendix I
Initial Contact to Hospitals/CDWs

To: …… Medical Center

From: ……., ARIC Coordinator

Copy: ……., ARIC Principal Investigator

Date: ……., 2016

Re: ARIC Study Request for Data in Electronic Format

The Atherosclerosis Risk in Communities (ARIC) Study is a National Institutes of Heart, Lung and Blood, NIH sponsored epidemiological study of the University of …… that has described trends in cardiovascular disease in …….. since 1987. Over the past three decades, your hospital has kindly provided a listing of discharge diagnoses to ARIC, from which ARIC sampled eligible hospital records for abstraction of selected items by our study personnel. These items are then used to classify hospitalized myocardial infarction (MI) and heart failure (HF) according to ARIC's study criteria (https://www2.cscc.unc.edu/aric/surveillance-manuals). An example ARIC publication on time trends in cardiovascular disease is *Rosamond WD et al., Trends in the incidence of myocardial infarction and in mortality due to coronary heart disease, 1987 to 1994. N Engl J Med. 1998 Sep 24;339(13):861-7.*

We are writing today because the NIH has asked ARIC to conduct a feasibility evaluation on the use of electronic health records (EHR) in health research. ARIC has already sampled and will abstract, in its usual fashion, a subset of records from the 2015 hospitalization index that your hospital provided. We would like to extract electronically the same data elements from EHR and compare them for accuracy and completeness to the information abstracted by the ARIC staff. To perform this evaluation, we are requesting an electronic copy of portions of the sampled hospitalization records. An overview of the data elements requested is included. These are the data elements that ARIC has abstracted from medical records for three decades.

To facilitate the task, we will provide a spreadsheet with an itemized listing of the structured data elements and portions of the text in the medical record to be extracted, based on commonly used data formats. To save time and avoid ambiguity, the spreadsheet will also specify items with reference to their location in the EHR clinical record.

As is the case for the data you have already been sharing with the ARIC study each year, these EHR data files will be uploaded to the ARIC Coordinating Center at the University of North Carolina at Chapel Hill using a secure file transfer protocol. The ARIC Coordinating Center will remove personal identifiers using the MITde-id scrubbing program, perform automated data extraction, encrypt the files you provided, and archive them on a secure server.

We look forward to discussing with you how to seek your hospital's approval, as well as who the contact for medical record extraction may be. ARIC, of course, will pay for any additional hospital resources needed to complete this EHR request.

Overview of the Data Elements Requested

I. The structured data elements we request are:

1. Admission and discharge dates
2. Demographic information, including age at admission, sex, race, weight, height
3. Vital Signs, including blood pressure, pulse (heart rate)
4. Laboratory values and standards, including LOINC codes, date stamps and time stamps
5. Medications including RxNorm codes, their date stamps and time stamps
6. Diagnosis codes for current diagnoses

The output file should include headings that label each type of data element

II. The unstructured (text) portions of the record we request are:

1. All physician notes
   - Discharge Summary
   - History and Physical
   - ED Notes
   - Cardiac Consultations
   - Other consultations
   - Progress Notes

2. All diagnostic exams
   - Chest X-Ray
   - Echocardiogram
   - Electrocardiogram (ECG)
   - Cardiac Catheterization
   - Cardiac Radionuclide Ventriculogram
   - Cardiac MRI
   - Cardiac CT
   - Cardiac Stress Test

We will provide a spreadsheet with requested output file formats. We suggest that a test dataset be prepared first (e.g., the first 10 records listed).

Since this is the first time ARIC has requested data in a standard electronic format, it would be helpful to first get your answers to the questions found on the following page by return e-mail or by calling the number provided. We will then deliver the list of records to be retrieved from your CDW or repository at your convenience, identified by hospital record number and date of discharge. This will be provided in a secure form, per your specifications. We will be glad to discuss this request with you and answer any questions. Please contact me by email at …….. or phone (....) … - ……...

Thank you for your continuing assistance with this important research.

[Hospital or Clinical Data Warehouse] _____

[Director or Contact]_____

ARIC Study Request for Data in Electronic Format

Kindly provide your answers to this brief questionnaire to ……., ARIC Coordinator, by e-mail at …… or by calling (…..) …-……

1. Does your hospital have a data analyst or "honest broker" (or informatics core/quality assurance/performance improvement team) that regularly creates datasets from your system's EHR system? ____Yes ____No

If yes:

2. Do data analysts have "back end"/direct access to EHR databases? Specifically, "business intelligence" level of access is not sufficient.

_____

3. From our list of requested data elements (above), please list which data elements your team is <u>not</u> capable of providing. What points of clarification are needed from our team?

_____

_____

An example of domains in PCORnet Common Data Model is on page 3 of http://pcornet.org/wp-content/uploads/2014/07/PCORnet_CDM_3_Lay_Guide_FINAL.pdf

We conservatively estimate that a data analyst/honest broker will require around 60 hours to complete this project.

Appendix II
Request for EHR Submitted to Hospitals/their CDWs

Appendix II.a – Cover memo

TO: Clinical Data Warehouse (CDW) Director
FROM: ARIC Surveillance Coordinator
COPY: ARIC Field Center PI
DATE: April xx, 2017
RE: ARIC EHR Feasibility Study – Retrieval of sections from the electronic health record

As previously mentioned, we are contacting you as part of the ARIC Electronic Health Records (EHR) Feasibility Study to request retrieval of portions of EHRs for selected hospitalizations which occurred at your hospital in the year 2015. A password-protected list of those hospitalizations, identified by the medical record number and date of discharge, is provided with this request.

To ensure that we have identified all relevant data element locations and that the process of data retrieval is as efficient as possible, we request an initial retrieval on the first five (5) records on the enclosed list of medical record numbers. Based on this initial record retrieval we will answer questions or modify our request if needed, before proceeding to retrieve the sections of EHRs for all listed medical record numbers.

Data Elements

We have identified all data elements to be extracted and provide those in the enclosed Excel workbook (ARIC_EHR_DataElements.xlsx). Data elements that we have identified as likely to exist in a structured format (not free text) within the CDW are listed in the first seven worksheets. The final worksheet (TEXT FILES) lists data elements from unstructured, free text/ "notes" data sources. Listed below are the eight worksheets and their descriptions. The *"Patient Encounter-to-data rows"* column in the table indicates if we expect to receive a single row of data per patient encounter (*one-to-one*) or if we expect that multiple rows of data may be outputted for a single patient encounter (*one-to-many*).

Worksheet Title	Type of Data	Patient Encounter-to-data rows	Description
DEMOGRAPHICS_ ENCOUNTER	Structured	One-to-one	Items from various relational tables within the CDW (domains) that can be combined into a single row for a given patient/encounter
CONDITIONS	Structured	One-to-many	Conditions reported by the patient or medical history/conditions that are carried to the current encounter.
DIAGNOSIS	Structured	One-to-many	Diagnoses during the encounter
PROCEDURES	Structured	One-to-many	Procedures recorded during the encounter
LABS	Structured	One-to-many	Laboratory measures
VITALS	Structured	One-to-many	Vital measurements, such as blood pressure, weight
MEDICATIONS	Structured	One-to-many	Medications
NOTES	Unstructured	One-to-many	List of text notes that are requested in addition to the structured data elements

Each worksheet lists the requested data elements by column. For each data element, we provide a description, the output value (if a specific mapping to categorical data is requested), and the output data format requested. Note that medical record number (MRN), discharge date (Discharge_Date), and the hospital-specific encounter ID (Encounter_ID) are listed in every worksheet. These are used for our study to link files. Examples of the requested data elements are shaded in green and are provided at the bottom of each worksheet.

For each data element, in each tab of the Excel spreadsheet, we provide information on the following attributes in the BLUE section of the sheet:

Data Element Attributes	Description/Instructions
FIELD LABEL	This is the name of the data element that we want in the output file. (e.g., the medical record number of the patient should be labeled "MRN")
ORDER	This is the order that this data element should appear in the output file.
LIKELY DOMAIN	This is where we think that a data element likely resides in most data warehouses.
DESCRIPTION	This is a description of the data element to assist you in identifying the corresponding data element (s) in your data warehouse.
VALUE_OUTPUT	For categorical structured data elements, these are the categories we expect to be available options for this data element in the data warehouse. If your data warehouse does not have a data element with these categories, or if in order to provide these categories you need to use multiple data elements, please provide that information in the pink section of the spreadsheet.
DATA_TYPE	This describes the required formatting for each data element. Below are the options (if *x* is not replaced with a value, any length is permissible): **Text(x):** Data element should be formatted as a text/character variable of length *x.* **Date(MM/DD/YYYY):** The data element should be formatted as a date variable in "DD/MM/YYYY" format. **Time (5) HH:MI:** Data element should be formatted as 24-hour clock time variable, with the first 2 integers representing the hour, followed by a ":", followed by 2 integers representing the minute. (e.g. "01:01", "14:01") **Integer(x):** Data element should be formatted as an integer with length *x.*

The CDW of your hospital may use different data element output values for categorical data or may require multiple variables to derive the values suggested by the "VALUE_OUTPUT" attribute. The PINK-shaded portion of each worksheet is included to give you an opportunity to provide us with feedback concerning such alternate organization of data elements.

Deliverables:

Please provide us with the following output:

1. Using the first 5 medical records numbers and their specified discharge date from the supplied list, please provide 8 pipe ("|") delimited text files.

a. Please name these files according to the worksheet tab names (e.g., “Demographics_Encounter.txt”, “Medications.txt”, etc.), such that each output file corresponds to a separate worksheet in the attached Excel spreadsheet (ARIC_EHR_DataElements.xlsx).

b. Please output data elements in the order of the **ORDER** attribute (as listed in the blue section of the spreadsheet) and assign column names specified by **FIELD LABEL** (also listed in the blue section of the spreadsheet). If your CDW does not have this data element to output, please still create an empty column and label it using the **FIELD LABEL** attribute.

c. Many rows of data may correspond to a single patient in most of the output files, such as medications or labs. We would like to be provided all rows of data that match to the medical record number and discharge date for the 5 patients.

d. To facilitate data retrieval we have provided, as a reference for you, synthetic data in the requested output format (pipe delimited text files). However, note that these are not intended to be complete or necessarily logical representations of the possible output values for each data type.

2. Please complete the PINK-shaded portion of each worksheet found in the attached Excel spreadsheet (ARIC_EHR_DataElements.xlsx), as applicable.

3. We greatly appreciate your help with this feasibility study. To facilitate its success, we would like to ask you to additionally provide a response to several questions that will allow us to understand more clearly the process of data retrieval from your Clinical Data Warehouse. Those questions are listed in a separate document titled “ARIC EHR Feasibility Study - Analyst Feedback” which is attached to this request. Please submit this via email to bbogle@email.unc.edu.

We will provide a secure portal for you to upload the requested output to the ARIC Coordinating Center (CC) at the University of North Carolina at Chapel Hill. If preferred, I can retrieve the records from you for a secure transfer to the ARIC CC. On arrival at the ARIC CC the records will be de-identified using the MITde-id software.

Please do not hesitate to contact me at < email address, phone > for any clarifications concerning this request.

Appendix II. b - Excel spreadsheet listing data elements requested, their format and location, and sample output record files

Available on request by contacting arichelp@unc.edu

Appendix III

Overview of Protocol for Processing Text Data into NLP Format (ver. 04.28.2017):

For each **HOSPITAL'S NOTES FILE:**

Step 1. Create Data File *S1_[HOSPID] from RECEIVEDDATA*

(SEE DATALINKS_STEP1 tab in *NLP_Process_Example.xlsx* document)

a. Assign HOSPID to each record using value corresponding to the hospital, from a table provided by the ARIC Coordinating Center
b. For each Patient/Encounter/Note, if multiple lines per note, combine lines into one text source.
c. Add *NOTE_TYPE_ID* based on NOTE_TYPE crosswalk
d. Add *NOTE_AUTHOR_TYPE_ID* based on AUTHOR_TYPE crosswalk
e. Add *NOTE_DEPT_ID* based on ENCOUNTER_DEPARTMENT crosswalk
f. For each record assign an ARIC ID using a medical record-ARIC ID crosswalk table provided by the ARIC Coordinating Center.
g. Remove the medical record column.

Step 2. Create data file S2*_[HOSPID]*: Run MIT DE-ID on *S1_[HOSPID]*

a. Use US Census/Birth Records to remove names from *NOTE_BODY_COMBINED*
b. Scrubbed Notes should be stored in variable called *NOTE_BODY_SCRUBBED*
c. Extract a sample for manual verification of de-identification

Step 3. Create data file S3*_[HOSPID]*: by pre-Processing *S2_[HOSPID]*

a. Remove special characters that are not compatible with CTAKES
(maintain outside file that is read into program during this task that is easily updated if we run into new issues)

b. Count the length of the number of characters (AFTER special character removal in part 3.a) in each note and store in "*Note_Length*" variable

Step 4. Create initial list of CTAKES concepts and negation terms as an easily editable table (csv file or other) that can be read into CTAKES script

DO UNTIL (QUALITY SUFFICIENT):

Step 5. Create data file *S4_[HOSPID]*: By running *CTAKES* on *S3_[HOSPID]*

Step 6. For each *CUI, c*reate a new variable *[CUI]_POS_N* and count occurrences for each note

Step 7. For each negation term/CUI combination, create a new variable *[CUI]_NEG_N* and count negation occurrences for each note.

Step 8. For each variable in Step 6 and Step 7, create a binary variable [*CUI]_Y and [CUI]_N .* For each note:

a. [*CUI]_Y =1 if [CUI]_POS_N>0 ; 0 otherwise*
b. [*CUI]_N =1 if [CUI]_NEG_N>0; 0 otherwise*

Step 9. Calculate relevant statistics and compare to gold standard

Step 10. Determine which negation terms and CUIs should be added to master list

Step 11. Add Step 10 terms to master list.

LOOP