American Journal of Preventive Cardiology 18 (2024) 100678

Contents lists available at ScienceDirect

American Journal of Preventive Cardiology

journal homepage: www.journals.elsevier.com/american-journal-of-preventive-cardiology

Houston Methodist cardiovascular learning health system (CVD-LHS) registry: Methods for development and implementation of an automated electronic medical record-based registry using an informatics framework approach

Khurram Nasir [a,b,*,**], Rakesh Gullapelli [b,*], Juan C Nicolas [b], Budhaditya Bose [b], Nwabunie Nwana [b], Sara Ayaz Butt [b], Izza Shahid [a], Miguel Cainzos-Achirica [c], Kershaw Patel [a], Arvind Bhimaraj [a], Zulqarnain Javed [b], Julia Andrieni [d], Sadeer Al-Kindi [a,b], Stephen L Jones [b], William A Zoghbi [a]

[a] Division of Cardiovascular Prevention and Wellness, Department of Cardiology, Houston Methodist DeBakey Heart & Vascular Center, Houston, TX, United States
[b] Center for Health Data Science & Analytics, Houston Methodist Research Institute, Houston TX, United States
[c] Hospital del Mar/Parc de Salut Mar and Barcelona Biomedical Research Park, Barcelona, Spain
[d] Population Health and Primary Care, Houston Methodist Hospital, Houston, TX, United States

Abbreviations and acronyms: ACE, Angiotensin-converting-enzyme; ACL, Adenosine Triphosphate-Citrate Lyase; ADI, Area Deprivation Index; AF, Atrial Fibrillation; AHA, American Heart Association; ALP, Alkaline Phosphatase; ALT, Alanine Aminotransferase; ARB, Angiotensin II Receptor Blockers; ARNI, Angiotensin Receptor-Neprilysin Inhibitor; ASCVD, Atherosclerotic Cardiovascular Disease; AST, Aspartate Aminotransferase; ATC, Anatomical Therapeutic Chemical Classification; BI, Business Intelligence; BMI, Body Mass Index; BNP, B-type Natriuretic Peptide; CAD, Coronary Artery Disease; CDC, Center for Disease Control and Prevention; CKD, Chronic Kidney Disease; CPD, Chronic obstructive pulmonary disease; CPT, Current Procedural Terminology; CRP, C-Reactive Protein; CVDLHS, Cardiovascular Disease Learning Health System; DBP, Diastolic Blood Pressure; DM, Diabetes Mellitus; ED, Emergency Department; EMR, Electronic Medical Record; ERD, Entity Relationship Diagram; ETL, Extract Transform Load; GLP1RA, Glucagon-like-Peptide 1 Receptor Agonists; HbA1c, Hemoglobin A1C; HDL, High-Density Lipoprotein; HMG-CoA, 3-hydroxy-3-methyl-glutaryl-coenzyme A; HIPAA, Health Insurance Portability and Accountability Act; HR, Heart Rate; IBD, Inflammatory Bowel Disease; ICD-10 CM, International Classification of Diseases, Tenth Revision, Clinical Modification; ICD-10 PCS, International Classification of Diseases, Tenth Revision, Procedure Coding System; IRB, Institutional Review Board; IT, Information Technology; PAD, Peripheral Artery Disease; PHI, Protected Health Information; LDL, Low-Density Lipoprotein; LHS, Learning health system; LPA, Lipoprotein A; MORTI, Methodist Online Research Technology Initiative; NAM, National Academy of Medicine; NHGIS, National Historical Geographic Information System; NIH, National Institute of Health; PCKS9, Proprotein convertase subtilisin/kexin type 9; RDB, Relational Database; SBP, Systolic Blood Pressure; SDOH, Social Determinants of Health; SGLT2, Sodium-Glucose Transport Protein 2; siRNA, Small or Short Interfering Ribonucleic Acid; SQL, Structured Query Language; SSIS, SQL Server Integration Services; SSMS, SQL Server Management Studio; SOP, Standard Operating Procedure; SVI, Social Vulnerability Index.

** Corresponding author at: Division of Cardiovascular Prevention and Wellness, Department of Cardiology, Houston Methodist DeBakey Heart & Vascular Center, 6550 Fannin St Suite 1801, Houston, TX 77030, United States.
E-mail address: knasir@houstonmethodist.org (K. Nasir).
* Co-first Authors

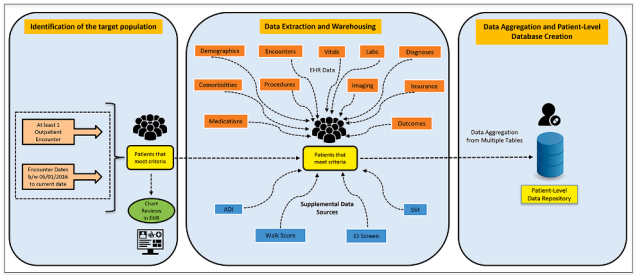
https://doi.org/10.1016/j.ajpc.2024.100678

Available online 30 April 2024

HIGHLIGHTS

- Big data applications provide comprehensive insights into care gaps.
- An EMR-based learning health system registry was developed to determine cardiovascular care gaps.
- Houston Methodist CVD-LHS registry includes longitudinal data of >1 million patients.
- CVD-LHS tracks incident and recurrent ASCVD.
- CVD-LHS identifies burden, determinants and at-risk patients for health management.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Patient registries
Population health
Cardiovascular research
Electronic medical records
Health outcomes-process assessment
Information retrieval-warehousing

ABSTRACT

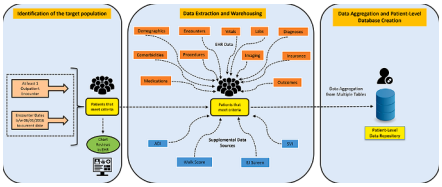
Objectives: To investigate the potential value and feasibility of creating a listing system-wide registry of patients with at-risk and established Atherosclerotic Cardiovascular Disease (ASCVD) within a large healthcare system using automated data extraction methods to systematically identify burden, determinants, and the spectrum of at-risk patients to inform population health management. Additionally, the Houston Methodist Cardiovascular Disease Learning Health System (HM CVD-LHS) registry intends to create high-quality data-driven analytical insights to assess, track, and promote cardiovascular research and care.

Methods: We conducted a retrospective multi-center, cohort analysis of adult patients who were seen in the outpatient settings of a large healthcare system between June 2016 - December 2022 to create an EMR-based registry. A common framework was developed to automatically extract clinical data from the EMR and then integrate it with the social determinants of health information retrieved from external sources. Microsoft's SQL Server Management Studio was used for creating multiple Extract-Transform-Load scripts and stored procedures for collecting, cleaning, storing, monitoring, reviewing, auto-updating, validating, and reporting the data based on the registry goals.

Results: A real-time, programmatically deidentified, auto-updated EMR-based HM CVD-LHS registry was developed with ~450 variables stored in multiple tables each containing information related to patient's demographics, encounters, diagnoses, vitals, labs, medication use, and comorbidities. Out of 1,171,768 adult individuals in the registry, 113,022 (9.6%) ASCVD patients were identified between June 2016 and December 2022 (mean age was 69.2 ± 12.2 years, with 55% Men and 15% Black individuals). Further, multi-level groupings of patients with laboratory test results and medication use have been analyzed for evaluating the outcomes of interest.

Conclusions: HM CVD-LHS registry database was developed successfully providing the listing registry of patients with established ASCVD and those at risk. This approach empowers knowledge inference and provides support for efforts to move away from manual patient chart abstraction by suggesting that a common registry framework with a concurrent design of data collection tools and reporting rapidly extracting useful structured clinical data from EMRs for creating patient or specialty population registries.

Central Illustration. Overview of the method development and implementation of an automated EMR-based patient registry



1. Introduction

Although substantial improvements in therapies for the treatment and prevention of cardiovascular disease (CVD) have been achieved over the past several decades, it remains the leading cause of morbidity and mortality in the United States. Despite efforts to identify and measure the disease burden, social determinants, persistent health disparities, and equitable healthcare gaps, the prevalence of CVD is projected to rise by 10% by 2030, posing a significant challenge to public health initiatives [1,2]. A better understanding of the real-world local evidence has not only the potential to exponentially support academic endeavors but also support population health initiatives.

The growing role of big data in healthcare presents an opportunity to address these challenges and improve health outcomes. The availability of vast amounts of data, including electronic medical records (EMRs), patient-reported outcomes, wearable devices, internet-derived data, and genomic information, offers a wealth of information for research and quality improvement purposes [3-5]. Harnessing this data has become critical for transforming cardiovascular care and establishing evidence-based research applications [6]. Today, EMRs serve as valuable data sources, capturing day-to-day patient care activities and generating a repository of aggregate data than traditional randomized controlled trials (RCTs) would allow [7-9]. This real-world evidence data can be used to validate cohorts, monitor patient outcomes in realtime, and improve the value and efficiency of healthcare [10-13].

With the widespread adoption of EMRs and the growing investment in real-world evidence and big data analytics by healthcare systems,

there exists an exceptional opportunity to develop and leverage a system-wide registry capable of providing comprehensive insights into care gaps and guiding optimal approaches for managing CVD and delivering the value in healthcare as described in Fig. 1. Such a registry could serve as the cornerstone of a comprehensive CVD learning health system (LHS), integrating knowledge generation at the core of clinical practice and care delivery.

In this paper, we present the framework for the Houston Methodist CVD Learning Health System (HM-CVD LHS) registry, a uniquely integrated research platform designed to investigate the real-world prevalence and determinants of atherosclerotic cardiovascular disease (ASCVD). The HM-CVD LHS registry aims to comprehensively study disease trajectories across the spectrum of at-risk patients, inform population health management strategies, and facilitate the implementation of risk mitigation interventions. By harnessing the wealth of data available through EMRs and implementing advanced analytics, the HM-CVD LHS registry has the potential to support a wide range of cardiovascular research domains, enhance clinical decision-making, and ultimately improve population health outcomes.

2. Methods

2.1. Study settings

This registry was developed at Houston Methodist (HM) Hospital in Texas, United States. The multi-center health system implements a single EMR system (Epic) across all inpatient and ambulatory settings in 8 locations including 27 cardiovascular specialty programs. Methodist Online Research Technology Initiative (MORTI) at HM Hospital determined that this research study is a non-interventional, retrospective cohort analysis, and thus a waiver of consent was approved by the institutional review board (IRB).

The study team designed and developed an automated electronic registry integrated with an EMR to identify, manage, and evaluate the patients diagnosed with established and at risk for ASCVD.

The study population consists of adult patients aged≥18 years that are presented or transferred to one of the hospitals in the health system with at least one encounter in outpatient settings between June 2016 and December 2022. De-identified clinical information on patient's demographics, vitals, diagnoses, laboratory tests & results, imaging tests & results, procedures, medications, comorbid conditions, clinical outcomes, International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10 CM) codes available in problems list, visit diagnosis and discharge diagnosis, the Current Procedural Terminology (CPT) codes listed in the professional billing transactions were obtained directly from the EMR for all visits and stored in the form of tables in the SQL server database. Key inclusion criteria and a list of key data elements collected for ASCVD patients and at-risk populations are described in Table 1. A standard operating procedure (SOP) manual was created to document all the methodological procedures including the snapshots of the key data elements which served as a user guide and training manual for all the stakeholders. The data dictionary and SOP facilitated transparency and communication between clinicians and analysts and served as an ongoing resource when questions arose about the data source or data elements. Several iterations of validation tests were conducted after every data refresh and when new data elements were added to the registry. Sufficient security and data privacy measures were implemented in terms of data collection, processing, storing, and maintaining the database as per the institutional policies and the regulations of the Health Insurance Portability and Accountability Act of 1996 (HIPAA).

2.2. Development of the CVD-LHS registry

2.2.1. Software tools

For this project, we employed our existing EMR, Epic's reporting data warehouse – Clarity, clinical documentation, reporting, and population health modules (all from Epic Systems, Verona WI). We also used our existing business intelligence (BI) tools such as Microsoft Visio (Version 2019, from Microsoft Corporation, Redmond, WA) to design an Entity Relationship Diagram (ERD) (as shown in **Supplemental Figure 1. Appendix A**). The ERD served as the fundamental concept of data modeling illustrating the logical relationships existing between multiple source tables in EPIC. Finally, SQL, a widely known programming language was used to pull the data at defined grains (**Central Illustration and** Fig. 2) from EMR into the relational database (RDB) model using the Extract-transform-load (ETL) process in Microsoft SQL Server Management Studio (SSMS) (Version 2019, from Microsoft Corporation, Redmond, WA). SQL Server instances provide efficient and systematic storage of a high volume of data with high performance, availability, scalability, flexibility, management, and security. We developed a constructive common registry framework using Transact SQL scripts and ETL scripts to pull the core data elements from Clarity into the SQL server database model as shown in Fig. 3.

The systematic process of building the CVDLHS registry using the common registry framework involved three key steps.

2.2.1.1. Identification of the target population (Disease classification). The schematic development of the HM CVD-LHS registry is illustrated in **Central Illustration**. The adult patient population included patients with or without established ASCVD. We reviewed existing clinical documentation of EMR's companion relational data warehouses "ICD Code Groupers" which contains several ICD-10 CM-related codes grouped under specific diagnoses and/or problems based on physicians'

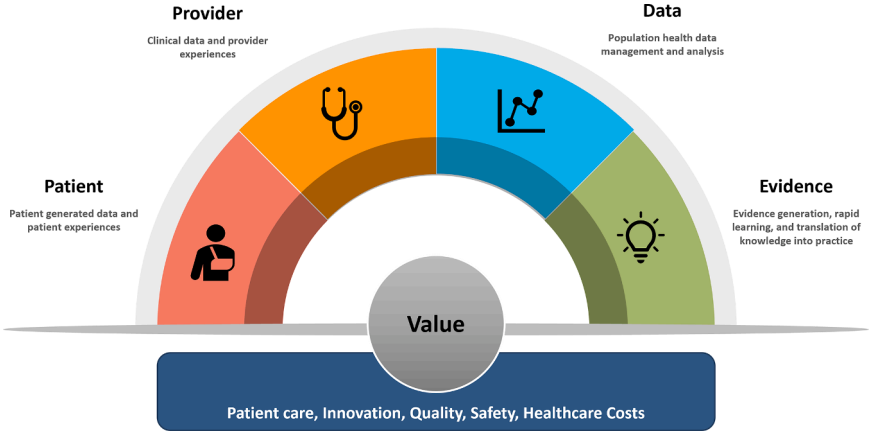


Fig. 1. Conceptual Framework of a Learning Health System.

Table 1
Inclusion Criteria and Key Data Elements in the CVDLHS Registry.

Inclusion Criteria	Only outpatient encounters (as shown in **Supplemental Table 1** in **Appendix A**) across the HM system (No hospital encounters). Encounter dates between 6/1/2016 and the current date. Having a visit diagnosis or problem list matching the expert-reviewed ICD-10 CM codes for CAD, PAD, and Stroke/ Cerebrovascular diseases (for ALL encounters). List of current medications for ALL encounters. List of specific labs (order procedures and associated components) for ALL encounters. No canceled appointments. Have the primary diagnosis code within the problem list at any time. Only valid patients (No test patients).
Key Variables	**Patients** Patient ID MRN Gender Race Ethnicity Age (at the encounter) Insurance/Payor **Encounters** Encounter ID Insurance type at encounter Encounter type Encounter Date Location of the Clinic List of Visit Diagnoses for the Encounter List of Problems for the Encounter Specialty/Department Total number of encounters and duration between the encounters **Vitals** Systolic Blood Pressure (SBP) Diastolic Blood Pressure (DBP) Heart Rate (HR) Weight Height Body Mass Index (BMI) **Labs** ***Lipid Panel*:** Total cholesterol, Low-Density Lipoprotein (LDL), High-Density Lipoprotein (HDL), and Triglycerides ***Blood Sugar*:** Glucose and Hemoglobin A1C (HbA1c) ***Other:*** Lipoprotein A - LP(a), Troponin, Creatinine, C-Reactive Protein, etc. **Medications** ***HMG-CoA Reductase Inhibitors:*** Atorvastatin, Cerivastatin Simvastatin, Lovastatin, etc. ***PCSK9 inhibitors:*** Evolocumab, Alirocumab ***Selective Cholesterol-Absorption Inhibitors:*** Ezetimibe ***Fibric Acid Derivatives:*** Clofibrate, Fenofibrate, Etofibrate, Gemfibrozil, etc. ***Bile Acid Sequestrants:*** Colestyramine, Colestipol, Colextran, etc. ***Adenosine Triphosphate-Citrate Lyase (ACL) Inhibitors:*** Bempedoic Acid ***Omega-3 Fatty Acids:*** Omega-3-Triglycerides Incl. Other Esters and Acids ***Small Interfering RNA (SIRNA):*** Inclisiran ***Miscellaneous Antihyperlipidemic Agents:*** Icosapent Ethyl, Cardiosterol, Probucol, etc. **Clinical Outcomes** Number of Emergency Department visits Number of Inpatient Visits Length of stay In-hospital mortality Admission and discharge diagnoses codes **Procedures/Imaging** Procedure or Imaging date Procedure or Imaging name CPT codes ICD-10 PCS codes

recommendations at the institutional level. The study team evaluated and validated each of the ICD-10 CM codes internally and externally by cardiologists within the Department of Cardiology, Houston Methodist DeBakey Heart and Vascular Center (Table 2). A comprehensive list of included and excluded ICD-10 CM Codes from the ICD Code Groupers was described in **Appendix B.**

The existing data model of the data warehouse employed standard EMR data in the form of source tables. ETL scripts were programmed to query the data warehouse and to retrieve the latest grain of data starting from June 2016 for a sample of patients (~50) having at least one outpatient encounter type (from the list provided in ***Supplemental Table 1*** in **Appendix A**) and having at least one of the diagnoses codes (from the listed ICD-10 CM codes in the ICD Code Groupers for coronary artery disease (CAD), peripheral artery disease (PAD), and Stroke/Cerebrovascular disease respectively in **Appendix B**) either from diagnosis or problem list tables. The sample patient medical record numbers (MRNs) were then used by cardiologists to review the patient charts in the EMR to confirm the diagnoses and other inclusion criteria. After this was done successfully, similar ETL tasks were then performed to retrieve all the patient MRNs meeting the study inclusion criteria as described in Table 1.

2.2.1.2. Data extraction and warehousing

2.2.1.2.1. EMR data. Standard EMR source tables which host the majority of clinical information and a core set of EMR structured tables were used for retrieving custom data by developing customized and replicable ETL scripts for each core variable such as diagnoses, problems list, encounters, orders, procedures, medications, comorbidities, etc. except for the flowsheet data in the SQL server database in a parallel fashion. SQL-stored procedures were used to perform post-processing the data and SSIS packages were developed to execute ETL tasks for auto-updating the process of data extraction and warehousing and were scheduled once a month for auto-refreshing the complete database.

We queried the encounter table in the EMR data warehouse after applying all the inclusion checklist criteria to retrieve the list of all the encounters followed by filtering the relevant outpatient encounter types (as listed in **Supplemental Table 1 in Appendix A**) and by matching the ICD-10 CM codes in the visit diagnosis and problems list tables with respective ICD-10 CM codes (as described in **Appendix B**). Duplicate patients were dropped to make a unique cohort of the patient population. (**Supplementary Table 2 in Appendix A**) All the relevant outpatient encounters and data (vitals, laboratory measures, diagnostic codes, medications, tagged ED, and hospitalizations as shown in **Supplementary Tables 3–13** in **Appendix A**) were extracted. In addition, Federal Information Processing Standards (FIPS) codes were used to map the census block group to classify patients based on various social/environmental validated indexes (Area deprivation index (ADI), social vulnerability index (SVI), neighborhood walkability score, environmental justice score) as shown in **Supplementary Tables 14–17** in **Appendix A)**

2.2.1.3. Data aggregation and patient-level database construction. Data aggregation is a multi-step process of combining all the clinical data of a single patient existing in one or more tables into one single repository or table at one row per patient grain. First, the process of data aggregation and consolidation started with building a SQL stored procedure where the query retrieved all the demographic and diagnostic information for each patient. After this step, the query extracted specific information about each patient at the first and last outpatient encounters from the respective logical tables created using the common registry framework and stored in the 'FINAL_OUTPUT' table. **Supplementary Table 18** includes the list of variables aggregated at the patient -level. In the next phase, the query aggregated laboratory data for the components listed in **Supplementary Table 19** for each patient at any encounter, at the first encounter, and at the last encounter. **Supplementary Table 20** includes the list of the medications used for this purpose. Further, the query aggregated the counts namely, the number of inpatient visits (hospitalizations) and ED encounters. **Supplementary Table 21** includes the counts for each patient. Lastly, the query linked the ADI and SVI data at the census tract level by geo-mapping the FIPS codes (as a linking variable) with the census block group codes available in the patient's most

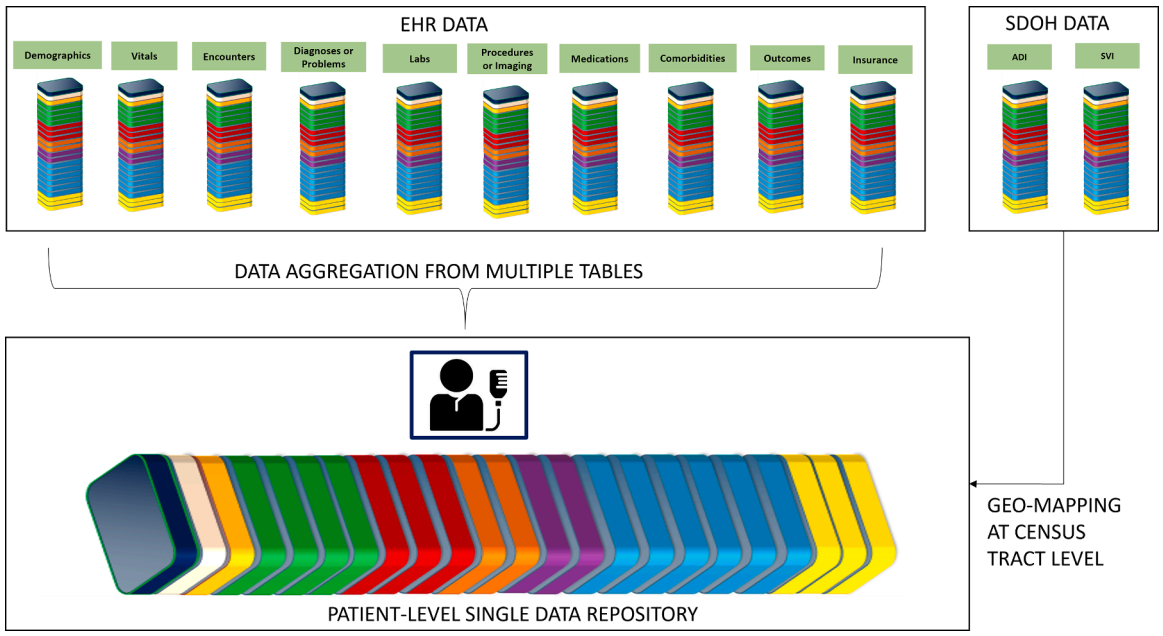


Fig. 2. An overview of data aggregation and patient-level data repository creation.

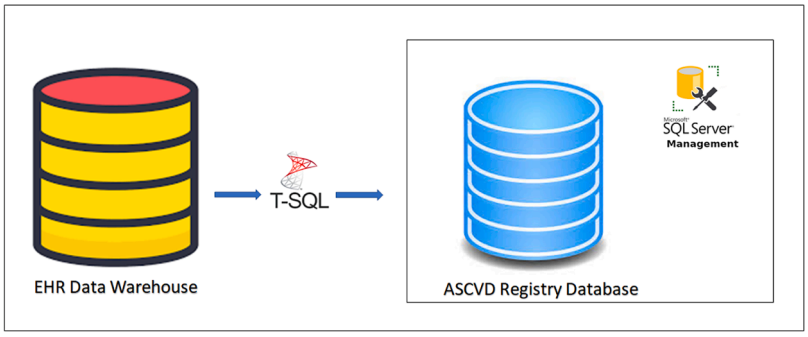


Fig. 3. Common Registry Framework.

Table 2
Definitions of events or outcomes in the registry.

Event/Outcome	ICD-10 CM Code/Definition
Coronary Artery Disease	I20 - I25
Peripheral Artery Disease	I70 - I73
Stroke/Cerebrovascular Disease	I60 – I69
Atherosclerotic Cardiovascular Disease	Any diagnosis of CAD, PAD, Stroke/ Cerebrovascular disease
Hypertension	I10 – I15
Obesity	BMI >30 kg/m^2
Diabetes Mellitus (Type I and II)	E10 – E11
Inflammatory Bowel Disease	K50 – K52
Chronic Kidney Disease	N18
Cancer	C00 – C97

recent residential address. Finally, the aggregated data is iteratively tested by querying the data for various epidemiological use cases and comparing the results with that of the extracted data from each of the individual logical tables by multiple analysts. To ensure validity of data, a retrospective chart review of a randomly selected cohort of 50 patients was conducted by two independent cardiologists within the team. The cardiologists examined various EMR components, particularly the visit diagnoses, admit and discharge diagnoses, and problem lists to verify at least two occurrences of the ICD-10 codes identifying the ASCVD population, thus minimizing misclassification. This process yielded a 100% accuracy rate, confirming the internal validity of our methodology. The validation of the data was completed successfully, and the stored procedure was scheduled for auto-update or auto-aggregation once a month with the same level of data i.e., one row per patient. An overview of the data aggregation process is illustrated in Fig. 2.

3. Results

The design, validation, and technical build of the registry occurred over the past two years. The registry was successfully built and implemented within the EMR data to meet the predefined specifications (variables and functionalities) as described in **Central Illustration**. The database consists of approximately 450 wide-range variables of interest with a few examples highlighted in Table 1.

3.1. Study population and baseline characteristics

3.1.1. Demographics

The CVD-LHS registry captured data for about 1,171,768 unique individuals; age≥18 years, median age 53 years (range, 18–108), 59% Women, 55% Non-Hispanic White (NHW) individuals, 14% Non-Hispanic Black (NHB) individuals, and 16% Hispanics, with or at risk of ASCVD who had at least 1 established outpatient encounter between June 2016 and December 2022 (Fig. 4). Table 3 summarizes baseline characteristics of the registry population across the spectrum of ASCVD (CAD, PAD, and Stroke/Cerebrovascular Disease) and demographics,

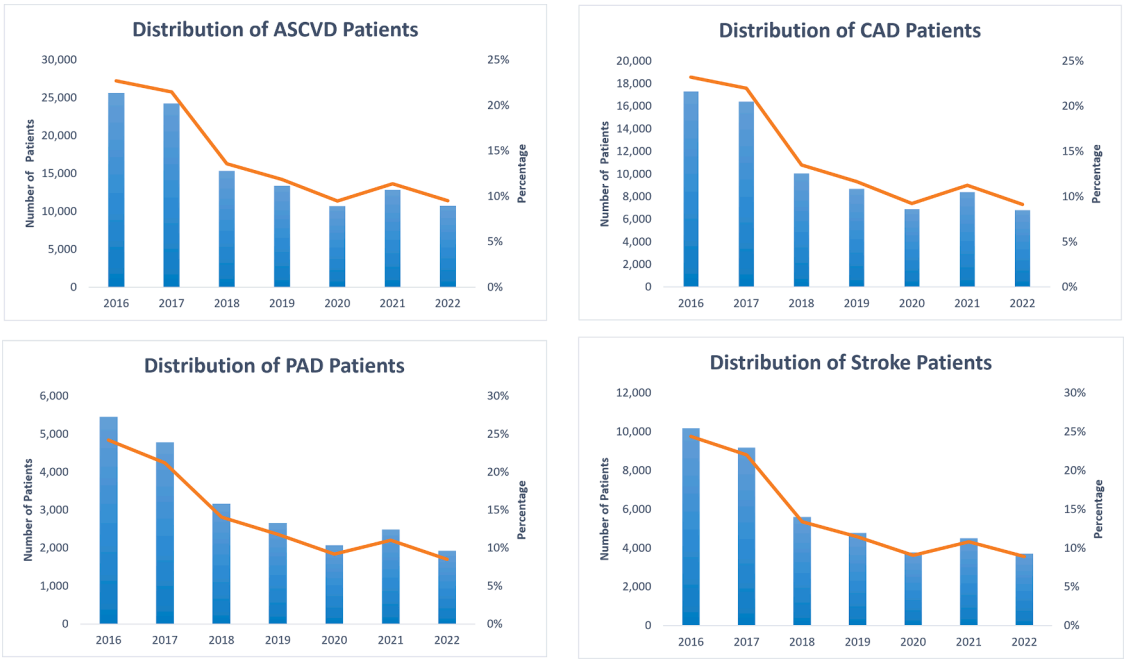


Fig. 4. Yearly distribution of total unique patients in the CVD-LHS registry.

vitals, ADI, labs, medications, and comorbid conditions.

3.1.2. *Outpatient visits, ED visits, and hospitalizations*

We found that, of all the participants in the CVD-LHS registry, a total of 8,130,730 outpatient encounters, 496,363 linked hospital admissions, and 633,991 emergency department encounters occurred between June 2016 and December 2022 across all the locations of the health system. Of the total population, 231,786 participants had at least one hospital admission while 293,965 participants had at least one emergency department visit during the study period. Overall, 18,517 adults (Men = 10,705; Women = 7812) had established MACE events recorded (Angina = 384; MI = 13,057; Stroke = 1391; Revascularization = 7550). The mean follow-up duration was 6 months from the first outpatient visit.

3.1.3. *Vitals*

When the latest data is considered for total participants, the mean (±sd) SBP and DBP were found to be 129.1 (±18) mmHg and 78.1 (±10) mmHg, respectively. Their mean (±sd) heart rate and BMI was found to be 77.1 (±13.7) bpm and 29.1 (±6.8) kg/m^2, respectively. Among those with BMI, much of the population (45%) was found to be having BMI >35 kg/m^2 (obese class 2+) followed by 41% with BMI >30 kg/m^2 (obese class 1+) and 14% with BMI >40 kg/m^2 (obese class3+).

3.1.4. *Laboratory*

At baseline, for those participants with recorded laboratory data available, the lipid panel results were found to be symmetric. 40% of the participants had LDL-C, HDL-C, and Triglycerides results. More than half of the participants have laboratory data available for creatinine (65%) and Fib-4 score (58%). About 40% had HbA1c, 15% had brain natriuretic peptide (BNP), 9% had c-reactive protein (CRP), and 0.6% lipoprotein A (LPA).

3.1.5. *ASCVD burden*

We identified 113,022 ASCVD patients (9.6%), ≥18 years of age at the time of diagnosis and having at least one diagnosis of CAD or PAD, or Stroke/Cerebrovascular Disease across all the encounters using the ICD-10 CM codes available in visit diagnosis and problems list tables in the EMR. Men (55%) were found to have a high prevalence of ASCVD when compared to women (45%). The older population (age ≥65) showed a high prevalence (>50%) of cardiovascular disease across the spectrum of ASCVD. NHW were found to have a high prevalence of ASCVD (64%) followed by NHB (15%), and Hispanics (12%). The unadjusted prevalence of CAD, PAD, and Stroke/Cerebrovascular disease is found to be 74,551 (6%), 22,537 (2%), and 41,755 (3.6%) respectively.

3.1.6. *Comorbidities*

At the time of listing the registry, almost 17% of the total participants had a history of diabetes mellitus, while other comorbid conditions are found to be prominent including hypertension (42%), cancer (9%), renal disease (8%), chronic obstructive pulmonary disease (COPD) (16%), inflammatory bowel disease (IBD) (1.2%) as shown in Table 3. The history of the same conditions when compared between ASCVD vs non-ASCVD were diabetes 43% vs 14%, hypertension 89% vs 37%, renal disease 31% vs 6%, COPD 31% vs 14%, IBD 1.6% vs 1.2%, cancer 18% vs 8%, respectively.

3.1.7. *Medications*

Overall statin use was found to be 27% in the total population as compared to 71% in those who were diagnosed with ASCVD and 22% in the non-ASCVD cohort. A similar trend was observed for high-dose statin use 9% in the total population vs 38% in ASCVD patients vs 6% in the non-ASCVD group. The prescription rates for any statin were found to be consistent in ASCVD patients across the spectrum including CAD (76%), PAD (69%), and Stroke (68%), and the rates for high-dose statins were found to be CAD (42%), PAD (34%), Stroke (35%). The medication utilization rates for non-lipid-lowering medications were shown in Table 3.

3.1.8. *Area deprivation index as measure of social vulnerability*

About 22% of our registry population is seen in lower ADI quintiles (least deprived) compared to 8% in higher ADI quintiles (most deprived). Interestingly, we found a higher prevalence of ASCVD (27%) in the ADI quintile 2 (second least deprived) than in the ADI quintile 5 (11%). (Most deprived). The same pattern was observed across the ASCVD spectrum, indicating that the ADI quintile 2 contains much of the population at risk.

Table 3
General Characteristics of Patients in the Registry.

	TOTAL	NONASCVD	ASCVD	CAD	PAD	STROKE
Overall Adults, N (%)	1171,768 (100)	1058,746 (90.4)	113,022 (9.6)	74,551 (6.4)	22,537 (1.9)	41,755 (3.6)
Sex, n (%)						
Male	484,767 (41.4)	422,599 (39.9)	62,168 (55)	45,634 (61.2)	12,076 (53.6)	19,237 (46.1)
Female	686,935 (58.6)	636,082 (60.1)	50,853 (45)	28,916 (38.8)	10,460 (46.4)	22,518 (53.9)
Age in years, mean (±sd)	52.3 ± 18.3	50.5 ± 17.9	69.2 ± 12.2	69.7 ± 11.5	71.1 ± 11.6	69.6 ± 12.9
Age Groups in years, n (%)						
18 - 39	333,856 (28.5)	331,721 (31.3)	2135 (1.9)	798 (1.1)	298 (1.3)	1116 (2.7)
40 - 64	487,665 (41.6)	455,263 (43)	32,402 (28.7)	20,833 (27.9)	5249 (23.3)	10,999 (26.3)
65 - 79	274,814 (23.5)	218,373 (20.6)	56,441 (49.9)	38,618 (51.8)	11,752 (52.1)	20,501 (49.1)
>=80	75,433 (6.4)	53,389 (5)	22,044 (19.5)	14,302 (19.2)	5238 (23.2)	9139 (21.9)
Race-Ethnicity, n (%)						
Non-Hispanic Whites	645,319 (55.1)	573,538 (54.2)	71,781 (63.5)	48,768 (65.4)	13,331 (59.2)	25,950 (62.1)
Non-Hispanic Blacks	162,372 (13.9)	145,312 (13.7)	17,060 (15.1)	9942 (13.3)	4425 (19.6)	6822 (16.3)
Non-Hispanic Asians	82,531 (7)	75,914 (7.2)	6617 (5.9)	4704 (6.3)	1019 (4.5)	2623 (6.3)
Non-Hispanic Other	43,176 (3.7)	40,913 (3.9)	2263 (2)	1535 (2.1)	396 (1.8)	727 (1.7)
Hispanics	186,045 (15.9)	172,681 (16.3)	13,364 (11.8)	8401 (11.3)	3036 (13.5)	4889 (11.7)
Unknown Ethnicity	52,325 (4.5)	50,388 (4.8)	1937 (1.7)	1201 (1.6)	330 (1.5)	744 (1.8)
Texas State ADI Quintiles, n (%)						
Quintile 1 (Least Deprived)	262,942 (22.4)	241,814 (22.8)	21,128 (18.7)	14,279 (19.2)	2994 (13.3)	8099 (19.4)
Quintile 2	333,796 (28.5)	303,369 (28.7)	30,427 (26.9)	20,513 (27.5)	5603 (24.9)	11,198 (26.8)
Quintile 3	279,249 (23.8)	252,317 (23.8)	26,932 (23.8)	17,762 (23.8)	5666 (25.1)	9762 (23.4)
Quintile 4	188,739 (16.1)	167,584 (15.8)	21,155 (18.7)	13,622 (18.3)	4907 (21.8)	7853 (18.8)
Quintile 5 (Most Deprived)	95,568 (8.2)	82,999 (7.8)	12,569 (11.1)	7853 (10.5)	3245 (14.4)	4554 (10.9)
ADI N/A	11,474 (1)	10,663 (1)	811 (0.7)	522 (0.7)	122 (0.5)	289 (0.7)
Blood Pressure, mmHg, mean (±sd)						
Systolic Blood Pressure	129.1 ± 18	128.5 ± 17.6	134.2 ± 20.6	133.8 ± 20.3	136.6 ± 22.8	134.4 ± 20.9
Diastolic Blood Pressure	78.1 ± 10	78.5 ± 9.9	75.3 ± 10.7	75.1 ± 10.6	73.9 ± 11	75.5 ± 10.7
BMI, mean (±sd)	29.1 ± 6.8	29.0 ± 6.9	29.1 ± 6.5	29.3 ± 6.4	28.6 ± 6.6	28.3 ± 6.3
Heart Rate, mean (±sd)	77.1 ± 13.7	77.5 ± 13.7	74.4 ± 13.6	73.7 ± 13.5	75.4 ± 13.8	74.8 ± 13.6
Obesity, n (%)	521,721	476,397	45,324	30,720	8592	15,210
Obese Class 1+	216,012 (41.4)	191,707 (40.2)	24,305 (53.6)	17,112 (55.7)	4528 (52.7)	8111 (53.3)
Obese Class 2+	232,932 (44.6)	218,378 (45.8)	14,554 (32.1)	9216 (30)	2774 (32.3)	5144 (33.8)
Obese Class 3+	72,777 (13.9)	66,312 (13.9)	6465 (14.3)	4392 (14.3)	1290 (15)	1955 (12.9)
Labs, n (%)						
LDL in mmHg	475,705 (40.6)	393,813 (37.2)	81,892 (72.5)	56,086 (75.2)	16,000 (71)	30,008 (71.9)
HDL in mmHg	476,712 (40.7)	394,944 (37.3)	81,768 (72.3)	56,001 (75.1)	15,974 (70.9)	29,942 (71.7)
TRIG in mmHg	476,047 (40.6)	393,483 (37.2)	82,564 (73.1)	56,521 (75.8)	16,178 (71.8)	30,198 (72.3)
Creatinine in mg/dL	764,188 (65.2)	661,838 (62.5)	102,350 (90.6)	68,123 (91.4)	20,653 (91.6)	37,420 (89.6)
Lipoprotein A in nmol/L	6599 (0.6)	3979 (0.4)	2620 (2.3)	2259 (3)	336 (1.5)	682 (1.6)
BNP in pg/mL	173,123 (14.8)	118,713 (11.2)	54,410 (48.1)	38,140 (51.2)	11,618 (51.6)	19,725 (47.2)
CRP in mg/dL	104,874 (9)	81,103 (7.7)	23,771 (21)	15,239 (20.4)	6055 (26.9)	9468 (22.7)
HBA1c in% of total Hgb	475,060 (40.5)	398,319 (37.6)	76,741 (67.9)	51,523 (69.1)	15,926 (70.7)	28,580 (68.4)
Fib4 Score	678,309 (57.9)	584,358 (55.2)	93,951 (83.1)	62,967 (84.5)	18,960 (84.1)	34,251 (82)
Medications, n (%)						
Statin	315,479 (26.9)	235,010 (22.2)	80,469 (71.2)	56,560 (75.9)	15,500 (68.8)	28,530 (68.3)
High Intensity Statin	107,706 (9.2)	65,069 (6.1)	42,637 (37.7)	31,594 (42.4)	7781 (34.5)	14,558 (34.9)
PCSK9	4682 (0.4)	2041 (0.2)	2641 (2.3)	2231 (3)	417 (1.9)	797 (1.9)
Ezetimibe	30,659 (2.6)	18,452 (1.7)	12,207 (10.8)	9711 (13)	2262 (10)	3982 (9.5)
Bempedoic Acid	685 (0.06)	441 (0.04)	244 (0.2)	200 (0.3)	53 (0.2)	86 (0.2)
Vascepa	9106 (0.8)	6290 (0.6)	2816 (2.5)	2272 (3)	628 (2.8)	819 (2)
Fibric Acid Derivatives	20,059 (1.7)	15,802 (1.5)	4257 (3.8)	3157 (4.2)	916 (4.1)	1341 (3.2)
Bile Acid Sequestrants	5178 (0.4)	4239 (0.4)	939 (0.8)	630 (0.8)	197 (0.9)	341 (0.8)
Cholesterol Absorption Inhibitors	23,930 (2)	14,901 (1.4)	9029 (8)	7229 (9.7)	1572 (7)	2839 (6.8)
Plavix	41,207 (3.5)	16,133 (1.5)	25,074 (22.2)	17,834 (23.9)	6447 (28.6)	9085 (21.8)
Brilinta	3971 (0.3)	1285 (0.1)	2686 (2.4)	2381 (3.2)	398 (1.8)	558 (1.3)
SGLT2	21,956 (1.9)	15,709 (1.5)	6247 (5.5)	4946 (6.6)	1355 (6)	1764 (4.2)
GLP1RAi	36,968 (3.2)	30,160 (2.8)	6808 (6)	4984 (6.7)	1505 (6.7)	2002 (4.8)
Aspirin	172,530 (14.7)	112,916 (10.7)	59,614 (52.7)	43,050 (57.7)	11,127 (49.4)	20,309 (48.6)
ARBs	99,232 (8.5)	79,345 (7.5)	19,887 (17.6)	14,186 (19)	3881 (17.2)	7052 (16.9)
ARNIs	6251 (0.5)	3114 (0.3)	3137 (2.8)	2734 (3.7)	568 (2.5)	746 (1.8)
ACE Inhibitors	120,504 (10.3)	98,257 (9.3)	22,247 (19.7)	14,918 (20)	4436 (19.7)	7728 (18.5)
Omega-3 Fatty Acids	70,956 (6.1)	57,053 (5.4)	13,903 (12.3)	10,228 (13.7)	2588 (11.5)	4716 (11.3)
Warfarin	9290 (0.8)	6309 (0.6)	2981 (2.6)	1999 (2.7)	717 (3.2)	1119 (2.7)
NOACs/DOACs	53,549 (4.6)	35,608 (3.4)	17,941 (15.9)	11,733 (15.7)	4251 (18.9)	7270 (17.4)
Comorbidities, n (%)						
Diabetes Mellitus	195,965 (16.7)	147,006 (13.9)	48,959 (43.3)	33,830 (45.4)	12,426 (55.1)	17,119 (41)
Inflammatory Bowel Disease	14,252 (1.2)	12,439 (1.2)	1813 (1.6)	1206 (1.6)	388 (1.7)	677 (1.6)
Hypertension	493,674 (42.1)	393,005 (37.1)	100,669 (89.1)	68,154 (91.4)	20,833 (92.4)	36,449 (87.3)
Renal Disease	95,350 (8.1)	60,523 (5.7)	34,827 (30.8)	24,168 (32.4)	9890 (43.9)	10,475 (25.1)
Cancer	104,515 (8.9)	84,429 (8)	20,086 (17.8)	13,813 (18.5)	4279 (19)	7354 (17.6)
Chronic Pulmonary Disease	186,814 (15.9)	152,171 (14.4)	34,643 (30.7)	24,115 (32.3)	8486 (37.7)	12,164 (29.1)

Footnote: Results are presented as either number (%) or mean (standard deviation).

4. Discussion

Cardiovascular disease (CVD) remains a leading cause of global morbidity and mortality, necessitating comprehensive population health initiatives and innovative research approaches. The availability of robust and integrated data is pivotal in effectively managing and preventing CVD, as it enables a deeper understanding of the disease and informs evidence-based clinical practice. Traditional patient registries have long served as invaluable resources in CVD research, providing insights into disease patterns, treatment outcomes, and population health trends. However, these registries often encounter challenges related to fragmented data sources, incomplete data capture, and labor-intensive manual processes, impeding their potential to drive impactful research and optimize clinical outcomes. Recent advancements in information technology and the widespread implementation of electronic medical records (EMRs) have revolutionized healthcare data management.

The HM CVD-LHS registry delineates an informatics-based registry framework, which incorporates existing EMR data to ascertain individual-level longitudinal information of patients with and spectrum of risk of ASCVD, with at least 1 outpatient encounter. The registry was able to accurately curate ASCVD data of >1 million patients from a large EMR system. In addition, the current registry also captures structured data of ~450 clinical variables directly from EMR, thereby facilitating the identification, burden, clinical and social predictors, and outcomes of ASCVD patients in a large integrated healthcare system. Designing an informatics-based common registry framework can be a complex endeavor, including the sequential review and construction of ICD-10 CM codes, definitions, data extraction, and warehousing frameworks and deploying them in EMR's warehouse and native SQL environment. However, pragmatic registries utilizing data obtained directly from EMR are often more practical than manual chart abstraction and amass a larger patient population [14]. Our study underscores the potential of an automatized in supporting evidence-based decision-making, offering a seamless flow of timely, comprehensive, and standardized data with the potential to enable in-depth investigations into CVD prevention, treatment, and outcomes as well as in supporting evidence-based decision--making and optimizing local clinical CVD prevention and management practices. This registry, which is one of the largest of its nature in the US to date confers several advantages.

First, the longitudinal nature of an individual person-level EMR data on ASCVD burden and risk enables the assessment of several important clinical and social surveillance variables which inform cardiovascular health, such as incident rates, reporting of adverse cardiovascular events, recurrent hospital admissions, the efficacy of medication, monitoring of individual-level risk factor trajectories, and appraisal of preventive strategies employed for high-risk patients [15]. The observed ASCVD prevalence of 9.6% in this CVD-LHS registry is comparable to the national prevalence of ASCVD in the US, which is suggested to be 8.0%. The minor variance may be attributed to several factors, such as age distribution, socioeconomic status, and underlying comorbidities. Additionally, the access to and utilization of healthcare services within our system might lead to more frequent diagnosis and reporting of ASCVD cases.

Second, the integration of social and clinical determinants can aid in identifying social vulnerabilities, care gaps, incidence, prevalence, and socioeconomic burden in conjunction with biological determinants which may be valuable to inform holistic and patient-centric decision-making [16]. Select socioeconomic, racial, and ethnic groups in the US have consistently demonstrated an increased burden of cardiovascular risk factors and adverse cardiovascular events [16]. Reasons for this are multifactorial and can largely be attributed to disparities in healthcare literacy, socioeconomic deprivation, implicit bias, and lack of access to optimal healthcare, all of which contribute to excess disparities in access and provision of cardiovascular care. To delineate the prevalence and influence of these factors in cardiovascular risk and disease in a community-based cohort, our registry includes a detailed appraisal of SVI and ADI. Patients in ADI quintile 2 had the highest prevalence of ASCVD compared with other ADI quintiles. This demonstrates that the ASCVD burden is not limited to the most socioeconomically vulnerable populations, but also extends to populations in the upper quintile.

Third, the detailed appraisal of this common registry framework can be reused for creating multiple patient registries across multiple specialties and patient sub-populations. Individual pipelines are being developed leading to the creation of additional patient registries such as diabetes mellitus, heart failure, etc. Ancillary grains of standard clinical data, irrespective of whether known to be associated with the current project, were also extracted, and stored in the database (Fig. 5) with the respective ETL frameworks to ensure that any future project needing the same grain of data could also be supported simultaneously. By integrating and analyzing data obtained from these registries, it would be possible to delineate the disease burden, biological risk factors, social determinants, medication utilization patterns, and clinical outcomes across diverse practices and populations. For example, ongoing individual pipelines are being developed leading to the creation of additional patient registries such as diabetes mellitus, heart failure, and other conditions.

Table	Granularity (one row per)	Patient	Date	Encounter / Encounter Type	Location/Specialty	Diagnosis	Medication	Outcomes	ICU
Outpatient Encounters	Encounter	X	X	X	X	X			
Hospital Encounters	Encounter	X	X	X	X	X		X	X
Flowsheet Data	Flowsheet Entry	X	X	X	X	X			
Imaging	Imaging Procedure	X	X	X	X	X			
Labs	Lab Result Entry	X	X	X	X	X			
Medications Administered	Administration	X	X	X	X	X	X		
Medications (Outpatient)	Medication	X	X	X	X	X	X		
Problems	Problem	X	X						
Procedures CPT	Procedure	X	X	X	X	X			
Procedures ICD-10	Procedure	X	X	X	X	X			
Social Hx	Social Item	X	X	X	X				
Derived Patient Aggregation	Patient	X	X			X	X		

Fig. 5. The granularity of the data in the CVDLHS Registry.

As with other EMR-based registries, this study has certain limitations. First, this is a single-center registry. Despite a large sample size, the geographical limitation of this registry suggests that the study population is not nationally representative. Second, although data obtained from EMR includes information across multiple domains, health profiles extracted from a single organization's EMR may contain missing information. Third, patients who initially reported at Houston Methodist outpatient center may have received follow-up or inpatient care elsewhere. This may cause some events to not be captured in the current database causing variable participation. Given these limitations, continuous efforts are needed to combine the registries with other robust data sources such as claims and administrative data to maximize their benefits. Fourth, current disease conditions are identified based on ICD-10 codes, however, for specific and targeted projects collaboration between data informatics and domain experts will be needed for 'phenotypic refinement' learning from prior experiences and validated algorithms developed by Phenotype Knowledge base and Health Dara Research UK Phenotype [17].

5. Conclusions

In summary, a new EMR-based automatized curated and harmonized CVD-LHS longitudinal real-world registry was designed and an interactive database with ~450 variables was developed successfully by extracting the clinical data for the patients with established ASCVD using their medical records. Designing the informatics-based common registry framework has proven to be a complex endeavor, including the sequential review and construction of ICD-10 CM codes, definitions, data extraction, and warehousing frameworks and deploying them in EMR's warehouse and native SQL environment. The registry enables earlier CVD diagnosis, creating longitudinal trajectories for the incident and recurrent ASCVD to study the impact of traditional and nontraditional factors on the presence/incidence of ASCVD, support for recruiting patients for clinical trials, developing EMR-based dramatic studies, and addressing quality of care projects with deeper real-time insights on current gaps and best practices, among other benefits. A learning health system model of this kind involves patients and their families partnering with clinicians and care teams, directly linked to a registry to support the health system's networks for outcomes improvement and research and offers an ideal framework for measuring what matters to a range of stakeholders interested in improving care for this special population. We believe the CVD-LHS registry has the potential to concord naturally with other registries of the institution bringing value to the health system and supporting patient-centered care, quality improvement, and scientific research.

Funding

No funding was utilized in the preparation of this manuscript.

CRediT authorship contribution statement

Khurram Nasir: Conceptualization, Data curation, Investigation, Methodology, Project administration, Supervision, Validation, Writing – review & editing, Resources. **Rakesh Gullapelli:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Visualization. **Juan C Nicolas:** Data curation, Formal analysis, Investigation, Methodology, Software, Validation. **Budhaditya Bose:** Data curation, Formal analysis, Investigation, Software, Validation. **Nwabunie Nwana:** Data curation, Formal analysis, Investigation, Software, Validation. **Sara Ayaz Butt:** Data curation, Formal analysis, Investigation. **Izza Shahid:** Investigation, Visualization, Writing – original draft, Writing – review & editing. **Miguel Cainzos-Achirica:** Data curation, Methodology, Project administration, Resources, Supervision, Validation. **Kershaw Patel:** Investigation, Methodology, Supervision, Writing – review & editing. **Arvind Bhimaraj:** Investigation, Methodology, Writing – review & editing. **Zulqarnain Javed:** Data curation, Investigation, Methodology, Project administration, Writing – review & editing. **Julia Andrieni:** Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **Sadeer Al-Kindi:** Investigation, Methodology, Project administration, Supervision, Writing – review & editing. **Stephen L Jones:** Conceptualization, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **William A Zoghbi:** Conceptualization, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Khurram Nasir reports a relationship with Nova Nordisk, Novartis, Esperion, Amgen, National Institutes of Health, and the Jerold B. Katz Academy of Translational Research that includes: consulting or advisory, funding grants, and speaking and lecture fees. All other authors report no relevant disclosures.

Acknowledgments

None.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ajpc.2024.100678.

References

[1] Mohebi R, et al. Cardiovascular disease projections in the United States based on the 2020 census estimates. J Am Coll Cardiol 2022;80(6):565–78.
[2] Prevention, C.f.D.C.a. Heart disease facts. 2023 [cited 2023 June 15]; Available from, https://www.cdc.gov/heartdisease/facts.htm.
[3] Lin H, et al. Using big data to improve cardiovascular care and outcomes in china: a protocol for the CHinese electronic health records research in yinzhou (CHERRY) study, 8. BMJ Open; 2018, e019698.
[4] Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system, 33. Health Affairs; 2014. p. 1163–70.
[5] Bates DW, et al. Big data in health care: using analytics to identify and manage high-risk and high-cost patients, 33. Health affairs; 2014. p. 1123–31.
[6] Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. Nat Rev Cardiol 2016;13(6):350–9.
[7] Gliklich RE, Leavy MB, Dreyer NA. Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User's Guide. Addendum 2 [Internet] 2019.
[8] Yang M, et al. Design and Implementation of a Depression Registry for Primary Care. Am J Med Qual 2019;34(1):59–66.
[9] Meltzer SN, Weintraub WS. The role of national registries in improving quality of care and outcomes for cardiovascular disease. Methodist Debakey Cardiovasc J 2020;16(3):205.
[10] Patel YR, et al. Development and validation of a heart failure with preserved ejection fraction cohort using electronic medical records. BMC Cardiovasc Disord 2018;18:1–8.
[11] Pike MM, et al. Improvement in cardiovascular risk prediction with electronic health records. J Cardiovasc Transl Res 2016;9:214–22.
[12] Floyd JS, et al. Validation of methods for assessing cardiovascular disease using electronic health data in a cohort of Veterans with diabetes. Pharmacoepidemiol Drug Saf 2016;25(4):467–71.
[13] Casey JA, et al. Using electronic health records for population health research: a review of methods and applications. Annu Rev Public Health 2016;37:61–81.
[14] Williams BA, et al. Establishing a national cardiovascular disease surveillance system in the United States using electronic health record data: Key strengths and limitations. Am Heart Assoc 2022.
[15] Vasan RS, Benjamin EJ. The future of cardiovascular epidemiology. Circulation 2016;133(25):2626–33.
[16] Havranek EP, et al. Social determinants of risk and outcomes for cardiovascular disease: a scientific statement from the. American Heart Association 2015;132(9): 873–98. Circulation.
[17] UK, H.D.R. Phenotypes. [cited 2023 July 12]; Available from: https://phenotypes.healthdatagateway.org/phenotypes/?