|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT452 | |
| **Submission Date** | 6/30/22 | |
| **Project Title** | An Algorithm for Ascertaining Coronary Heart Disease Cases and Controls from the Electronic Health Record | |
| **Tentative Lead Investigator** *(first author)* | Alborz Sherafati1, Seyedmohammad Saadatagah1 | |
| **Tentative Lead Investigator Email Address** | Sherafati.alborz@mayo.edu | |
| **Tentative Senior Author**  *(last author)* | Iftikhar J. Kullo1 | |
| **eMERGE Site Sponsor & Contact** | Mayo Clinic | |
| **All Other Authors** | Marwan Hamed1, Lubna Alhalabi1, Ozan Dikilitas1, … | |
| **Sites Participating** | 1 Mayo Clinic  2 Vanderbilt  3 Harvard  4 UAB  5 Mount Sinai  6 University of Washington  7 Columbia  8 Northwestern University | |
| **Background / Significance** | Coronary heart disease (CHD) is the leading cause of death in the United States1. The prevalence of CHD among adults ≥20 years in the US is 7.2%1. We propose the development and validation of electronic phenotyping algorithms to identify CHD and controls from the EHR.  We previously used the following algorithm to ascertain CHD: at least two related diagnostic codes (ICD-9 and ICD-10) on separate occasions within a 5-day window, or at least one relevant procedural code for coronary revascularization in the EHR2. This algorithm has high sensitivity and specificity, but the positive predictive value is less than ideal at 80%. Our goal is to refine this algorithm to improve the PPV while maintaining high sensitivity and specificity.  **CHD algorithms in other studies:**   * Ivers et al. proposed an algorithm based on searching keywords like ischemic heart disease, angina, and myocardial infarction in medical history fields. The sensitivity of their algorithm was 74.7%, but the specificity was 99.3%3. * Floyd et al. validated an algorithm for CHD, which included ICD-9 codes for previous MI, angina, and previous revascularization. The code for previous MI had a low sensitivity of 38%, but the codes for chronic conditions had higher sensitivity4. * Gronski et al. used an algorithm for myocardial infarction which was based on an abnormal troponin level and diagnosis of myocardial infarction in medical notes. This algorithm had a sensitivity of 100%5.   **References:**   1. Virani SS, Alonso A, Aparicio HJ, Benjamin EJ, Bittencourt MS, Callaway CW, et al. Heart Disease and Stroke Statistics-2021 Update: A Report From the American Heart Association. Circulation. 2021;143(8):e254-e743. 2. Dikilitas O, Schaid DJ, Kosel ML, Carroll RJ, Chute CG, Denny JC, et al. Predictive utility of polygenic risk scores for coronary heart disease in three major racial and ethnic groups. The American Journal of Human Genetics. 2020;106(5):707-16 3. Ivers N, Pylypenko B, Tu K. Identifying patients with ischemic heart disease in an electronic medical record. J Prim Care Community Health. 2011;2(1):49-53. 4. Floyd JS, Blondon M, Moore KP, Boyko EJ, Smith NL. 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. 5. Gronski L, Martinson W, Singh KP, Ryan J. Utility of daily troponin orders for identifying acute myocardial infarction patients for quality improvement. Crit Pathw Cardiol. 2012;11(2):74-6. 6. Bangash H, Pencille L, Gundelach JH, Makkawy A, Sutton J, Makkawy L, et al. An implementation science framework to develop a clinical decision support tool for familial hypercholesterolemia. Journal of personalized medicine. 2020;10(3):67. 7. Sohn S, Ye Z, Liu H, Chute CG, Kullo IJ. Identifying abdominal aortic aneurysm cases and controls using natural language processing of radiology reports. AMIA summits on translational science proceedings. 2013;2013:249. 8. Savova GK, Fan J, Ye Z, Murphy SP, Zheng J, Chute CG, et al., editors. Discovering peripheral arterial disease cases from radiology notes using natural language processing. AMIA Annual Symposium Proceedings; 2010: American Medical Informatics Association.   … | |
| **Outline of Project** | **Methodology:**  We will validate an updated CHD algorithm and compare its performance with three other CHD algorithms. Once validated, the best performing algorithm can be uploaded to the PheKB and used in genome-wide association studies.  The target population is adult (>18 years old) participants in research studies representing the general population. The expected prevalence of CHD in EHR cohorts is 7-10%. The four CHD algorithms represented in Table 1 will be run to detect CHD among the participants. Manual chart review will be performed as a gold standard to confirm the presence or absence of CHD. The data from the algorithms and chart reviews will be pooled for analysis.  Controls will be defined as participants with no procedural or diagnostic codes for CHD and no evidence of myocardial ischemia, myocardial infarction, and revascularization on chart reviews.   |  |  |  | | --- | --- | --- | | **Table 1.** CHD Algorithms | | | | **Algorithm** | **Description** | **Performance in Mayo Clinic participants** | | I | 1 procedural code or  2 diagnostic codes (hard events) at least 5 days apart | Sensitivity: 100%  Specificity: 98%  NPV: 100%  PPV: 83% | | II | 1 procedural code or  2 diagnostic codes (hard events) within 5 days | Sensitivity: 100%  Specificity: 97%  NPV: 100%  PPV: 80% | | III (combined) | 1 procedural code or​  2 diagnostic codes (hard) at least 5 days apart​  AND 2 diagnostic codes (hard) within 5 days | Sensitivity: 100%  Specificity: 98%  NPV: 100%  PPV: 88% | | IV | 1 procedural code or  3diagnostic codes | Sensitivity: 100%  Specificity: 98%  NPV: 100%  PPV: 88% | | |
| **Desired Data - Common Variables\***  *(Available from the CC)* | Demographics  ICD9/10 codes  CPT codes  Phecodes  BMI | Common Variable Labs  Common Variable Meds  ☐ Geocoding 2015 ACS variables  Other: Case/Control status |
| **Other Desired Data *(Available from participating sites)*** | *Please specifically list out any data elements that participating sites would collect or extract from clinical or other sources for this project (i.e. not common variables above)* | |
| **Desired Genetic Data** | eMERGE I-III Merged set (HRC imputed, GWAS)  eMERGE PGx/PGRNseq data set  eMERGEseq data set (Phase III)  eMERGE Whole Genome sequencing data set  eMERGE Exome chip data set  eMERGE Whole Exome sequencing data set  Other (not listed above): | |
| **Does project pertain to an existing eMERGE Phenotype?** | Yes, if so please list: **the CHD phenotype on PheKB**  No | |
| **Planned Statistical Analyses** |  | |
| **Ethical Considerations** |  | |
| **Available Funding or Resources** |  | |
| **Target Journal** |  | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | July 30, 2022: Determine validating site(s) specific chart review plan / initial data pull.  September 15, 2022: Discuss issues with initial chart review and running the algorithms.  October 30, 2022: Deadline to submit chart reviews and algorithm results (can be extended).  November 30, 2022: Data analysis and manuscript draft complete. | |

**\*Common Variables available across all datasets:**

* Demographics: sex, year of birth, decade of birth, race, ethnicity
* Codes: (repeated values & age at event): ICD, CPT, Phecodes
* BMI: (repeated value & age at event) height, weight, BMI
* Labs: (lab name, repeated lab value & age at event) Serum total cholesterol, LDL, HDL, Triglycerides, Glucose fasting/non-fasting/unknown, & White Blood Cell count
* Medications: (medication name, repeated, & age at event) Cerivastatin sodium, Rosuvastatin, Simvastatin, Fluvastatin, Pravastatin, Lovastatin, Atorvastatin, & Pitavastatin
* Other: Case/Control status on Phase I and Phase II phenotype: only on GWAS dataset participants