|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT457 | |
| **Submission Date** | 7/28/22 | |
| **Project Title** | Hidden Socio-technical gaps in centralized data solutions for multisite research network: experiences from the eMERGE IV project | |
| **Tentative Lead Investigator** *(first author)* | Cong Liu | |
| **Tentative Lead Investigator Email Address** | [cl3720@cumc.columbia.edu](mailto:cl3720@cumc.columbia.edu) | |
| **Tentative Senior Author**  *(last author)* | Chunhua Weng | |
| **eMERGE Site Sponsor & Contact** | Columbia University Irving Medical Center | |
| **All Other Authors** | George Hripcsak, Wendy Chung, Krzysztof Kiryluk, Casey Ta, Atlas Khan, Dean Karavite, John Connolly, Erin Cash, Emma Perez, Fei Wang, Sabrina Suckiel, Sarah Bland, Kate Mittendorf, Luke Rasmussen, Firas Wehbe, Alexandra Miller, Hana Bangash, Robert Freimuth | |
| **Sites Participating** | CUIMC, CHOP, CCHMC, MGB, Sinai, VUMC, NW, Mayo, (UAB, UW) | |
| **Background / Significance** | The eMERGE IV research program aims to recruit 25,000 diverse participants from ten health systems across the United States, order two types of genetic tests, collect clinical risk factors and family health history, and return the genome-informed risk results for ten conditions to participants and their providers. Representing the currently typical modern design for centralized data collection and analysis, a REDCap-based system (i.e. R4) is provided to all eMERGE IV sites for data submission. However, the inevitable idiosyncrasies in workflows among the ten clinical sites necessitate the development of site-specific IT infrastructures to satisfy the local study requirements and comply with local regulations. This paper summarizes the heterogeneous site-specific approaches employed to implement IT infrastructures to support the eMERGE IV research, the identified technical gaps and unmet user needs revolving the current R4 design, the tradeoffs between centralized and distributed data networks, and the lessons learned for designing and implementing IT infrastructures for supporting future large-scale multicenter clinical studies. | |
| **Outline of Project** | This research engages multidisciplinary stakeholders to investigate the sociotechnical gaps and to understand the unmet stakeholder user needs.  Phase I, the paper will conduct a survey (the survey instrument to be developed) to review what IT infrastructures have been implemented locally during the first two years of the eMERGE IV study in order to provide support for (1) data exchange with the centralized R4 system; (2) local workflow customization for recruitment and retention of participants; (3) monitor biospecimen collection and genetic test order; (4) clinical data extraction for Genomic Informed Risk Assessment (GIRA) reports generation; and (5) return of the GIRA reports to participants and their providers.  Phase II, the paper will interview site coordinators to understand reasons why this local IT infrastructure was implemented. The tentative reasons including (1) local regulatory requirement; (2) local study tracking requirements (with a focus on why the centralized platform cannot meet the study tracking requirement); (3) data or sharing with existing local research infrastructures; (4) requirement to optimize the recruitment and return of the results workflow. Other reasons will be explored during the interview. In addition, during the interview, the challenges and resources required in implementing such local IT infrastructures will be investigated.  Phase III, the paper will interview site developers and IT leaders to understand the reusability and generalizability of the IT components developed by each site. For developers, interview questions will focus on technique feasibility including whether the standardization was used, tech-stack dependency, code modularization as well as documentation level. For IT leaders, interview questions will focus on understanding the aspects of regulatory (e.g. site IP policy prohibit sharing) and resource management (e.g. budget limitation for developing a reusable tool).  Finally, the paper will summarize the lessons learned from the current IT implementation of eMERGE 4 and provide suggestions to improve IT infrastructure implementation for future similarly multicenter clinical studies. | |
| **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)*  Description of local IT implementation activities will vary. | |
| **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  No | |
| **Planned Statistical Analyses** | Most likely descriptive statistics only | |
| **Ethical Considerations** | None | |
| **Available Funding or Resources** |  | |
| **Target Journal** | Journal of Biomedical Informatics; Journal Of Implementation Science | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | 8/21/2022: Deadline for sites to submit interview answers. Phase I Result discussion and interview planning in the EHRI meeting.  9/11/2022: Finalizing interview for developers and coordinators. Phase II/III Results discussion.  10/02/2022: Finalizing interview for IT leaders. Phase III Results discussion and start manuscript drafting  11/05/2022: Draft circulated among writing group  11/30/2022: Final manuscript submitted for review | |

**\*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