|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT328 | |
| **Submission Date** | 2/5/19 | |
| **Project Title** | Genomic Data: Building a Path from Lab to Clinic | |
| **Tentative Lead Investigator** *(first author)* | Nephi Walton | |
| **Tentative Senior Author**  *(last author)* | Marc Williams | |
| **All Other Authors** | Darren Johnson, Luke Rasmussen, Robert Freimuth, Casey Overby | |
| **Sites Participating** | Geisinger, Northwestern, John Hopkins, Mayo | |
| **Background / Significance** | The role of genetic data in precision medicine has been talked about for years. Only recently have we had the ability to actually store discreet genetic data elements within the electronic health record (EHR) with two major EHR vendors now providing dedicated areas to store this type of data. While these advancements show promise in meeting some of the goals of precision medicine, there is little information about how the data should get to the medical record and which information should be stored in discreet form.  By creating a new place in the EHR designed for this data we have created what might be thought of as the genetic problem list, however automatically storing all reported variants in this format becomes a problem itself. We have stored variants of uncertain significance in PDF’s in the medical record for more than a decade but there has justifiable concern raised when this same information is suggested for implementation in a discreet form. When this type of information becomes front and center on a “genetic problem list” there are concerns that arise about the validity and certainty of the information presented with worries that information might be taken at face value and acted on inappropriately by providers who might not be as experienced with genetic information. Solid definitions or criteria for variants to enter the EHR in discreet form are needed to avoid manual entry of this information and assure the validity and certainty of the information that is transmitted through an automated process.  Another problem outside of sending the variant itself is assigning a disease or some other type of meaning to a variant. While in pharmacogenomics there are well defined variant and indication relationships, the majority of variants for disease that enter the system are novel and will not have predefined relationships to disease. Data structures (HL7/FHIR) for transmitting variants electronically do not currently enable the passing of indication with the variant. This is despite the fact that these variants are associated with disease on the test reports that come back from the lab. Some have suggested the manual entry of such variants however this does not allow for scale as genetic data begins to play a larger role in healthcare.  We propose to perform an analysis the current state of genetic data standards and their current use in EHR’s. We will define five use cases and apply them using current data standards and identify weaknesses in each use case. We will also propose automated mechanisms for vetting variants for entry into the EHR in each case, and automated methods for assignment of indication to variant for each case.  Use cases will consist of:   1. Two pharmacogenomic variants. 2. One disease where only one gene is thought to cause disease. 3. One disease with a genetic basis with an associated disease modifying variant. 4. One disease where multiple genes cause the disease. 5. One gene where variants in the same gene cause different diseases.   Based on these cases we will make recommendations for methods of automating the electronic transfer of discrete data into the EHR and define recommendations for selecting data that should be stored discreetly in the EHR. | |
| **Outline of Project** | 1. Review of current data structure of discrete genetic data elements in EHR 2. Compare with current standards (HL7, VCF) 3. Outline 5 clinical use cases for genomic data from eMERGE dataset. 4. Design filtering methods for each use case 5. Evaluate data structures for transfer (based on current standards) 6. Outline algorithms for assigning indication to variants 7. Cumulative analysis of results 8. Develop recommendations | |
| **Desired Data - Common Variables\***  *(Available from the CC)* | Demographics  ICD9/10 codes  CPT codes  Phecodes  BMI | Common Variable Labs  Common Variable Meds  Other: Case/Control status on Phase I and Phase II phenotypes |
| **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  No | |
| **Planned Statistical Analyses** | Qualitative analysis of current data structure and workflows. | |
| **Ethical Considerations** | None | |
| **Target Journal** | JAMIA, JAMIA Open | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | * March 2019 - Definition and assignment of clinical cases for analysis * April 2019 - Case analyses due * May 2019 - First Draft * June 2019 Review of Second Draft * July 2019 Submit article | |

**\*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