|  |  |  |
| --- | --- | --- |
| **eMERGE Network: Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT369 | |
| **Submission Date** | 11/27/2019 | |
| **Project Title** | Complex Survey Designs under Sampling Frame Misclassification | |
| **Tentative Lead Investigator** *(first author)* | Aya Mitani | |
| **Tentative Senior Author**  *(last author)* | Jonathan Schildcrout | |
| **All Other Authors** | Sebastien Haneuse and Nathaniel Mercaldo | |
| **Sites Participating** | VUMC & Harvard | |
| **Background / Significance** | Research for populations derived from electronic health records (EHR) is ubiquitous. However, it is well known that EHR data are challenged by missing and mis-measured values even in standard, demographic fields. This project seeks to explore the impact of mis-measured EHR data when the goal is to conduct a survey for a population of patients derived from an EHR. Oversampling relatively rare subgroups of patients (e.g., minorities) was crucial to the CERC-eEMERGE substudy design because such minorities are under-represented in medical and public health research and they are particularly under-represented in research involving perspectives on biobank research. EHR data were used to derive the sampling frame for the CERC survey at VUMC (and all other sites) and a disproprortionate stratified sampling design was conducted on a cross-classification of six, potentially mismeasured demographic variables. This study seeks to understand, from a methodological perspective, the impact that misclassification of variables used to define the sampling frame can have on the validity and efficiency of these studies. The data from the CERC survey will appropriately be used as a motivating example since it was the driving force behind this line of statistical research. | |
| **Outline of Project** | We will investigate the effects of utilizing an imperfect sampling frame on the design and analysis of complex survey data. This study is motivated by a large multi- center survey developed to elicit perspectives on biobank participation among under-studied populations (e.g., racial and ethnic monitories). A stratified sampling scheme was implemented to enrich the sample with low prevalence populations using a sampling frame primarily constructed from electronic health record data (EHR). Incomplete EHR data were imputed using geocode-derived census summaries which resulted in a well-defined, but imperfect sampling frame. We will show that under stratum misclassification: 1) complex study designs still result in more diverse samples compared to random sampling, 2) the efficiency of design-based estimators change as a function of the relative size of the sampling strata, and 3) analytic methods that account for the design are still required for valid inferences. We explore the effects of stratum misclassification in a real-world example by analyzing a subset of the biobank survey data from a single site, Vanderbilt University Medical Center. | |
| **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  ☒N/A |
| **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)*  **CERC survey data from VUMC only** | |
| **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):  ☒N/A | |
| **Does project pertain to an existing eMERGE Phenotype?** | Yes, if so please list  No | |
| **Planned Statistical Analyses** | This paper will involve design and model-based analyses of simulation and CERC data. Pertaining to the CERC data, we will examine the relationship between patient demographics (e.g., race, ethnicity, gender, age, rural/urban, etc.) and study outcomes that will include perspectives on biobank research, trust in the healthcare system, and trust in clinicians. Simulations will mimic the CERC data analysis but will impose varying degrees of misclassification on the sampling variables. | |
| **Ethical Considerations** | NA | |
| **Target Journal** | Applied biostatistics journal: Statistical Methods in Medical Research or Statistics in Medicine. | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | 12/15/2019: Approval  12/31/2019: Data preparation  6/30/2020: Methods developed and simulation studies completed  9/30/2020: CERC study analysis completed  12/31/2020: Manuscript submitted  12/31/2021: Project completed | |

**\*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