

**DEVELOPMENT AND PERFORMANCE OF TEXT-MINING ALGORITHMS TO  
EXTRACT SOCIOECONOMIC STATUS FROM DE-IDENTIFIED ELECTRONIC  
HEALTH RECORDS**

BRITTANY M. HOLLISTER

*Vanderbilt Genetics Institute, Vanderbilt University, 519 Light Hall, 2215 Garland Ave. South  
Nashville, TN, 37232, USA*

*Email: [Brittany.M.Hollister@Vanderbilt.edu](mailto:Brittany.M.Hollister@Vanderbilt.edu)*

NICOLE A. RESTREPO

*Institute for Computational Biology, Department of Epidemiology and Biostatistics, Case Western Reserve University,  
Wolstein Research Building, 2103 Cornell Road, Cleveland, OH 44106, USA*

*Email: [nrestrepo@case.edu](mailto:nrestrepo@case.edu)*

ERIC FARBER-EGER

*Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center, 2525 West End  
Avenue, Suite 600, Nashville, TN 37203, UA*

*Email: [eric.h.farber-eger@vanderbilt.edu](mailto:eric.h.farber-eger@vanderbilt.edu)*

DANA C. CRAWFORD

*Institute for Computational Biology, Department of Epidemiology and Biostatistics, Case Western Reserve University,  
Wolstein Research Building, 2103 Cornell Road, Suite 2527, Cleveland, OH 44106, USA*

*Email: [dana.crawford@case.edu](mailto:dana.crawford@case.edu)*

MELINDA C. ALDRICH<sup>†</sup>

*Department of Thoracic Surgery and Division of Epidemiology, Vanderbilt University Medical Center, 1313 21<sup>st</sup>  
Avenue South, 609 Oxford House, Nashville, TN 37232, USA*

*Email: [melinda.aldrich@vanderbilt.edu](mailto:melinda.aldrich@vanderbilt.edu)*

AMY NON<sup>†</sup>

*Department of Anthropology, University of California, San Diego, 9500 Gilman Drive #0532  
La Jolla, CA 92093, USA*

*Email: [alnon@ucsd.edu](mailto:alnon@ucsd.edu)*

---

<sup>†</sup> Co-Senior authors

Socioeconomic status (SES) is a fundamental contributor to health, and a key factor underlying racial disparities in disease. However, SES data are rarely included in genetic studies due in part to the difficulty of collecting these data when studies were not originally designed for that purpose. The emergence of large clinic-based biobanks linked to electronic health records (EHRs) provides research access to large patient populations with longitudinal phenotype data captured in structured fields as billing codes, procedure codes, and prescriptions. SES data however, are often not explicitly recorded in structured fields, but rather recorded in the free text of clinical notes and communications. The content and completeness of these data vary widely by practitioner. To enable gene-environment studies that consider SES as an exposure, we sought to extract SES variables from racial/ethnic minority adult patients (n=9,977) in BioVU, the Vanderbilt University Medical Center biorepository linked to de-identified EHRs. We developed several measures of SES using information available within the de-identified EHR, including broad categories of occupation, education, insurance status, and homelessness. Two hundred patients were randomly selected for manual review to develop a set of seven algorithms for extracting SES information from de-identified EHRs. The algorithms consist of 15 categories of information, with 830 unique search terms. SES data extracted from manual review of 50 randomly selected records were compared to data produced by the algorithm, resulting in positive predictive values of 80.0% (education), 85.4% (occupation), 87.5% (unemployment), 63.6% (retirement), 23.1% (uninsured), 81.8% (Medicaid), and 33.3% (homelessness), suggesting some categories of SES data are easier to extract in this EHR than others. The SES data extraction approach developed here will enable future EHR-based genetic studies to integrate SES information into statistical analyses. Ultimately, incorporation of measures of SES into genetic studies will help elucidate the impact of the social environment on disease risk and outcomes.

## 1. Introduction

### 1.1. *Socioeconomic status and health*

Socioeconomic status (SES) is a major determinant of variation in health outcomes worldwide<sup>1</sup>. SES is typically defined as a combination of income or wealth, educational achievement, and occupation<sup>2,3</sup> and can be assessed at the individual, household, or neighborhood level. Health outcomes within the United States, ranging from cancer to hypertension, vary by socioeconomic levels, regardless of how they are measured<sup>4</sup>. Multiple measures of SES have been previously associated with health outcomes, including income<sup>5</sup>, years of education<sup>6,7</sup>, occupational prestige<sup>2,8,9</sup>, insurance coverage<sup>10</sup>, and homelessness<sup>11</sup>.

SES likely affects health through various pathways including access to healthcare services, knowledge of health behaviors, exposure to environmental stressors and hazards, limited financial resources, and social support<sup>2</sup>. The relationship between SES and health is also highly entangled with race/ethnicity, as SES may covary with race and contribute in part to the existence of racial disparities in health<sup>4,12</sup>. Though these pathways are difficult to distinguish and could affect different populations to varying degrees, it is important to consider SES as a representation of these potential pathways in studies of human health.

Despite the overwhelming evidence that SES affects health outcomes, SES measures are often not included in genetic studies of disease. Neglect of SES data may be due to a lack of available SES information in existing cohorts, as well as the additional time and resources it takes to collect SES data in new studies. Despite these difficulties, it is vital to include measurements of SES in genetic association studies of racial disparities in health. In addition to the possible confounding that may occur due to the association of race/ethnicity with both SES and health<sup>13</sup>, SES has the potential to modify the effect of genetic variants on health outcomes<sup>14</sup>. Therefore, the etiology of disease is likely to be misunderstood without the inclusion of SES data in association studies. Although prior genetic association studies have found some gene variants that may explain a small

portion of racial disparities in disease prevalence and risk<sup>15</sup>, SES factors are likely to play an even larger role in racial health disparities<sup>6,7</sup>.

### **1.2. SES data within electronic health records**

The use of electronic health records (EHRs) for research purposes is becoming increasingly prevalent. With the announcement of the Precision Medicine Initiative and its goal of recruiting one million participants with biological and EHR data, the research use of EHRs is anticipated to increase<sup>16</sup>. EHRs provide an attractive resource for biomedical researchers for many reasons, including their rich phenotypic and longitudinal data, as well as the lower cost of participant recruitment versus a traditional prospective cohort study. Additionally, clinical biobanks that contain biological samples linked to EHRs are becoming an invaluable resource for conducting genetic epidemiology studies. Despite the potential for EHRs in research settings, these clinical data repositories currently have noted deficits in the availability and completeness of important social and environmental data<sup>17</sup>, including SES, that are known to contribute independently to health status and could modify genetic effects<sup>18</sup>.

Recognizing the importance of formally and consistently capturing social and behavioral measures in the EHR, the Institute of Medicine (IOM) recently recommended SES measures, specifically educational attainment, financial resource strain, and neighborhood median household income be included in the EHR<sup>19</sup>. The committee also recommended that a plan be developed by the NIH to expand the research use of EHRs to include social and behavioral data<sup>19</sup>. Adoption of these recommendations will take time, and may not be universal across medical centers; therefore, there is a need to develop approaches and methods to access existing unstructured SES data within the EHR for research purposes. SES data are almost entirely found within the free text clinical notes from providers and the clinical communications between providers. Currently, there are no published algorithms available to extract SES data from EHRs. In this study, we developed an approach for extracting available SES information from the free text of a de-identified EHR. These algorithms will facilitate the immediate extraction of key SES information from clinical biobanks for incorporation into future biomedical research.

### **1.3. BioVU**

BioVU is a DNA biobank of the Vanderbilt University Medical Center (VUMC) linked to de-identified EHRs. DNA samples are extracted from discarded blood samples drawn for routine clinical care<sup>20</sup>. DNA samples are linked to the Synthetic Derivative (SD), the de-identified version of the VUMC EHR, by a unique study ID. Medical records within the SD are scrubbed of all HIPAA identifiers such as names, locations, zip codes, and social security numbers. Dates within each SD record are shifted to prevent re-identification of the records. Date shifting is consistent within a single patient's record. As previously described<sup>21</sup>, data from BioVU are de-identified in accordance with provisions of Title 45, Code of Federal Regulations, part 46 (45 CFT 46); consequently, this study is considered non-human subjects research by the Vanderbilt University Institutional Review Board.

## 2. Methods

### 2.1. Population

The study population included all racial/ethnic minority patients >18 years old participating in BioVU as of 2011<sup>22</sup>. The EHRs used for the development of the algorithms were updated in 2015 to include current information. Race/ethnicity is administratively reported in BioVU and strongly correlated with genetic ancestry<sup>23,24</sup>. The majority (81%) of patients in the dataset are black individuals with an average age of 50 years (Table 1). The mean number of clinic visits within a patient's EHR record is 40.45 visits, and the mean number of days between patients' first and last visit within the EHR is 2,340 days (Table 1).

Table 1. Vanderbilt BioVU racial/ethnic minority population characteristics

Characteristic	n= 9,977
Sex	
Male	3,568 (36%)
Female	6,409 (64%)
Race/ethnicity	
Black	8,078 (81%)
Hispanic	1,049 (10.5%)
Asian	850 (8.5%)
Age (mean, years $\pm$ SD)	49.8 $\pm$ 18.1
Number of clinic visits (mean $\pm$ SD)	40.5 $\pm$ 55.0
Number of days between visits (mean $\pm$ SD)	2,340 $\pm$ 1,793.1

### 2.2. Development of algorithms

We sought to develop algorithms to extract SES data from structured and unstructured data in the de-identified EHRs. We developed seven algorithms for the extraction of SES information including education level, occupation, unemployment, retirement, insurance status, Medicaid status, and homelessness (Table 2). The initial development of the SES algorithms began with a manual review of both structured and unstructured data within the de-identified EHR of 200 randomly selected patients within this minority population dataset to identify the following: 1) the categories of SES information most frequently mentioned, 2) where in the EHR this information is noted, and 3) the semantic language used by clinical providers for socioeconomic information (Figure 1). The manual review revealed that the SES data were found exclusively within the unstructured free text of the clinical notes, social history, and clinical communications of this EHR. It was also noted that the most frequently mentioned semantic categories were employment, education, insurance status, and homelessness, and thus these categories were chosen for extraction. Semantic tags for each category were selected if they appeared more than once within the 200 development records.

#### 2.2.1. Employment

Employment information was extracted using three different algorithms designed to capture data on occupation, unemployment, and retirement. The occupation algorithm extracts the occupation

Table 2. Variables extracted by socioeconomic status (SES) algorithms applied to de-identified electronic health records

Semantic category	Format of algorithm output
Occupational prestige	0-100
Unemployment	Ever/never
Retirement	Ever/never
Education	-Never attended -Less than high school -High school graduate/GED -Associate's degree -Bachelor's degree -Master's degree -Professional degree -Doctoral degree
Uninsured	Ever/never
Medicaid	Ever/ never
Homelessness	Ever/never

mentioned in a patient's record and translates it to an occupational prestige score (scale of 0-100). This score represents how well-respected an occupation is within a society (i.e., subjective socioeconomic position). Occupational prestige scores were determined from a National Opinion Research Center (NORC) survey where respondents were asked to rank occupations according to their prestige<sup>25</sup>. The occupation tags utilized for the occupation algorithm were adopted from the most recent NORC report. The algorithm's occupation tags were shortened to 678 occupations from the original NORC list of 860 occupations given that some of the occupations were highly specific with repetitive occupational prestige scores. As an example, "teacher, elementary school" and "teacher, secondary school" were collapsed to "teacher."

We next used the occupation algorithm to search the unstructured data of the original 200 patients for the initial occupation tags. This search identified a large number of false positives, where the algorithm tagged occupation-related words that were not indicative of the patient's occupation. In order to filter these false positives, additional prefix language such as "works as," "is a/n," "employed" was added to a subset of occupations to filter out non-relevant terms, which greatly improved the algorithm. Unemployment data were extracted using semantic tags for unemployment (e.g., "unemployed," "does not work," "hasn't worked since"). The unemployment algorithm was then tested on the unstructured data from the 200 records used for development, and a high number of false positives were returned. These false positives were often in reference to medications. Therefore the tags "if this does not work" and "if that does not work" were excluded to filter false positives. Unemployment was classified as ever/never (Table 2). Retirement was also extracted from the EHR using the tag "retired" and classified as ever/never (Table 2).

### 2.2.2. Education

The education algorithm was designed to assign education level to a patient based on the highest education achieved and recorded in the EHR. Education levels were assigned to each relevant tag

word or phrase found in the unstructured text of the EHR (Table 2). Sixty-two semantic tags were utilized and the highest level of education was determined for each patient. These tags were exclusive to an assigned education level. For example, the high school degree category of education level included tags such as “high school graduate” and “completed 12<sup>th</sup> grade,” while the bachelor’s degree category included terms such as “BS degree” and “completed college.” The levels of education were based on U.S. census definitions with one modification such that all grade levels below high school graduate were collapsed into a “less than high school” category. We searched through the unstructured text of the 200 records used for development to determine if further filtering or modification was needed. Fifteen additional tags were used to filter false positive results related to types of medical education (e.g. “diet education,” “dialysis education”) and Vanderbilt Medical School students (e.g., “medical student,” “pharmacy student,” “student nurse”).

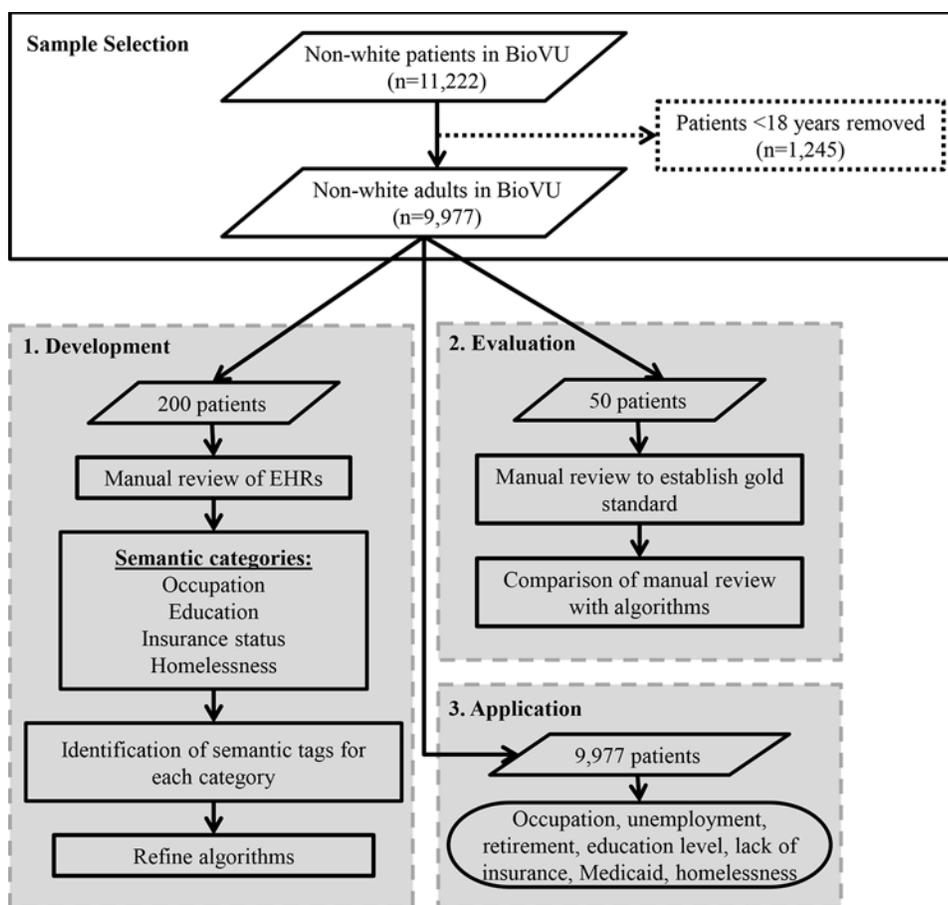


Figure 1. Overview of the development process for the SES algorithms

### 2.2.3. Insurance status

The extraction process for insurance status required two algorithms. The first algorithm was used to determine if there was any time point in the EHR when the patient did not have insurance based on the presence of five semantic tags (Table 2). These tags included “no insurance” and “does not have insurance” and excluded some language that was used in a standard discharge letter and

therefore appeared frequently in the EHR. A second insurance algorithm extracted Medicaid information using specific phrases or keywords such as “Medicaid” and “TennCare” and was classified as ever/never in order to determine if a patient was ever on Medicaid in their EHR (Table 2).

#### 2.2.4. Homelessness

Homelessness information was extracted using the tags “homeless” and “shelter” among the 200 development EHRs. After this search, several false positives were returned relating to patients who worked or volunteered at homeless shelters. Therefore, exclusion tags were added such as “volunteer at homeless shelter,” “works at homeless shelter,” “works with homeless,” and “animal shelter.” Homelessness was classified as ever/never (Table 2).

### 2.3. Evaluation of algorithm performance

To evaluate the performance of these SES algorithms, results were compared to findings from a manual review of 50 randomly selected patients. These 50 individuals were selected using random sampling without replacement. Two independent reviewers manually reviewed the clinical record of each patient and any discrepancies were resolved by discussion between the two reviewers. Comparison of results from the two independent reviewers was quantified using percent positive agreement, percent negative agreement, and kappa statistics for each of the seven categories and subcategories: education level, occupation, unemployment, retirement, uninsured, Medicaid, and homelessness. The manual review of 50 records was then compared to the algorithm results for each of the seven categories and subcategories. Sensitivity, specificity, and positive predictive value were estimated. The chi-square statistic was used to determine if the algorithms performed differently in different populations.

## 3. Results

### 3.1. Population characteristics

Among the total study population (n=9,977), we were able to extract at least one type of SES information from 8,282 (83.0%) individuals. We extracted education information for 3,780 individuals and occupation information for 7,296 individuals (Table 3). For the remaining

Table 3. Percent of records within the study population with algorithm-identified SES characteristics

Characteristics	Race			Total (n=9,977)
	Black (n=8,078)	Hispanic (n=1,049)	Asian (n=850)	
% with occupation	76.0	57.1	65.4	73.1
% unemployed	21.4	13.0	13.4	19.8
% retired	19.8	4.9	11.2	17.5
% with education	39.1	28.7	37.9	37.9
% uninsured	19.5	15.6	11.5	18.4
% on Medicaid	20.5	13.9	7.9	18.7
% homeless	3.7	1.3	1.0	3.2

categories, we were able to determine if an individual was unemployed, retired, uninsured, on Medicaid, or homeless at any point in his or her record. Of the total population for which we were able to extract SES data (n=8,282), 1,978 individuals were unemployed, 1,742 individuals were retired, 1,839 individuals were uninsured, 1,865 were on Medicaid, and 318 were homeless at least one time in their EHR (Table 3). For each of the seven categories, the algorithms returned SES information for a higher percentage of black patients than Hispanic or Asian patients ( $p < 0.00001$ ).

The five most frequently extracted occupations among those having occupation information (n=7,296) were manager, nurse, Army, manufacturer, and restaurant employee. Within the population with education information (n=3,780), the vast majority of individuals had a high school degree (n=2,066), followed by individuals without a high school degree (n=492), and individuals with a bachelor's degree (n=446).

### 3.2. Algorithm Performance

Prior to evaluating algorithm performance, the manual review results from the randomly selected records of 50 patients were compared between the two reviewers and any conflicts were resolved. The percent positive agreement between reviewers ranged from 98.0% to 100.0% and the percent negative agreement ranged from 94.7% to 100.0%. The Kappa statistic between reviewers ranged from 0.94 to 1.0.

Table 4. Comparison of manual review with algorithm results for each SES algorithm in a subset of randomly selected individuals (n=50)

Semantic Category	Records with SES information (%)	Sensitivity (%)	Specificity (%)	PPV (%)
Education level	48.0	66.7	84.5	80.0
Occupation	80.0	87.5	40.0	85.4
Unemployment	40.0	70.0	93.3	87.5
Retirement	14.0	100.0	90.7	63.6
Uninsured	8.00	75.0	78.3	23.1
Medicaid	18.0	100.0	95.1	81.8
Homelessness	2.00	100.0	95.9	33.3

Once all reviewer discrepancies were resolved, the manual review results were used as the gold standard and compared to the algorithm results. All the algorithms, with the exception of occupation, had very high specificity levels  $>78\%$ . The lower specificity for occupation (40%) is due to six of the ten individuals who did not have occupation information (as identified by manual review) but were identified as having occupation information by the algorithm. All the algorithms also had high sensitivity levels (above 70%), with the exception of education level (66.7%) (Table 4). The lower sensitivity for education is driven by eight individuals who have an education level that was identified by manual review but not by the algorithm. The lower sensitivity for unemployment is due to the six individuals who were identified as unemployed by manual review but not by the algorithm. PPV values across the algorithms ranged from 23.1%-87.5%. The lower PPV for the retirement algorithm (63.6%) is due to the four individuals identified as retired by the

algorithm but not retired by the manual review. (Table 4). The low PPV for the uninsured algorithm (23.1%) is due to the ten individuals who were identified as uninsured by the algorithm, but not by manual review. The low PPV for homelessness (33.3%) was a result of the fact that the manual review only identified one patient with homelessness in their record, whereas the algorithm misidentified two others.

### 3.2.1. *Missing data*

Of the total population (n=9,977), the algorithm was not able to extract any SES information for 1,695 individuals (17.0%). Of this group, there were 1,193 blacks, 309 Hispanics, and 193 Asians. Missing SES data were more common among Hispanic and Asian individuals, than among black individuals ( $p < 0.001$ ). The Hispanic and Asian populations represent 10.5% and 8.5% of the total dataset, respectively; however, these groups represent 18.2% and 11.4%, respectively, of the individuals with missing SES data. Males represent 35.8% of the study population and 28.0% of those without extracted SES data. The mean age for the total population is 49.9 years, and the mean age for the group without extracted SES information is 46.7 years.

## 4. Conclusion

Socioeconomic status is considered a fundamental cause of disease, because it affects so many proximate risk factors and disease outcomes<sup>26</sup>. SES has been consistently associated with health outcomes such as mortality, cancer, and cardiovascular disease<sup>27,28</sup>. Despite these consistent associations, SES data are typically not included in genetic studies of health outcomes. For studies that utilize biobanks, the lack of SES data is likely related to the difficulty in accessing these data within the EHR, where they are not usually recorded in structured fields. The algorithms described in this study are the first to extract these important data from EHRs for research purposes.

The SES algorithms described here focus on the extraction of data related to four semantic categories: occupation, education, insurance status, and homelessness. The occupation algorithms extracted and classified data as occupational prestige, unemployment (ever/never), and retirement (ever/never). The occupational prestige algorithm had a strong sensitivity and PPV; however it had a low specificity of 40%, reflective of the difficulty in filtering the occupation information. Although we took steps to remove false positives, it was difficult to completely eliminate all false positives without removing a large amount of accurate data. Our unemployment and retirement algorithms had high sensitivity (70% and 100%) and specificity (93.3% and 90.7%). While the unemployment algorithm had a high PPV, the retirement algorithm had a low PPV. Both unemployment and retirement were classified as ever/never because the EHR only captures a snapshot of time when the patient visits the clinic. It was not possible to accurately capture the length of time for unemployment or retirement as the patient's visits to the clinic may not reflect the length of time he or she was unemployed or retired. The sensitivity of the unemployment algorithm was affected by the varying language used to describe unemployment, which was identified in manual review but not consistently recognized by the algorithm ("does not work outside the home", "used to work in a restaurant"). The quality of the retirement algorithm was

affected by false positives related to the identification of words related to retirement that were used in a context outside of the patient's retirement from an occupation.

The education algorithm identified the highest level of education that a patient achieved over the course of their EHR. This algorithm had a high specificity and PPV, but a low sensitivity. The low sensitivity was due to the inability of the algorithm to detect variations in education level compared with the manual review. The variation in language used by clinical providers made it difficult to include every mention of education while still maintaining some level of precision. For example, some of the Vanderbilt Medical School students were excluded ("medical student," "pharmacy student") because of the frequent mention of these terms in the EHR related to patient care, rather than education level. The reviewers were able to infer education level based on occupation and context clues as well as identify the medical school students, while the algorithm was not able to do so. The algorithm that identified patients who were uninsured at some point in his or her record as well as the homelessness algorithm each had high sensitivity and specificity, but low PPV. Uninsured patients are the smallest portion of patients within VUMC, making up only 4.7% of the patient population in 2015<sup>29</sup>. The low PPV of these algorithms may be due to a low prevalence of uninsured patients and homeless individuals within the VUMC patient population. Within our randomly selected minority patient population used for evaluation, only four individuals were uninsured and one was homeless. These categories had the lowest prevalence within our evaluation dataset. The Medicaid algorithm was the highest performing algorithm, with a high sensitivity, specificity, and PPV.

The major challenges in utilizing EHR data in a research setting include missing data and the inconsistencies in the recording of SES data by clinical providers. While the majority of individuals within the study population had some SES information, a notable percentage of individuals did not have any SES information within their records (17.0%). The missing SES data could be a result of the lack of recording of information by the provider, either due to SES factors not being discussed in conversation with the patient, a low number of visits in the patient's EHR, or the willingness of the patient to provide SES information. Additionally, when variables are missing within a patient's record, we cannot distinguish whether it is due to negative data or just missing data. For example, if a patient does not have an occupation listed, we cannot assume that they are unemployed because it may have not been discussed with the provider or recorded by the provider. The higher level of missing data observed for Hispanic and Asian individuals in this dataset could be a reflection of the fact that the algorithms are optimized for the largest racial/ethnic population within the dataset (i.e., black patients).

The inconsistencies in the recording of the SES data are typical for social and environmental exposure data contained within free clinical text<sup>17</sup>. In the development of these algorithms, we noted that providers, in general, do not follow patterns when recording SES data within their notes in the EHR. The lack of consistent language and the numerous variations used to describe the SES information made extracting this information challenging. Furthermore, algorithms could also be limited by the accuracy of the selected filters and tags, rather than the information available within the EHR. While the aim of the algorithms was to include all possible semantic tags, there is a possibility that some information was missed by the algorithms or that information was captured inaccurately due to the limitations of the filtering process.

In addition to these general limitations, the algorithms developed here have specific limitations regarding portability. Even within the same dataset, we have noted a difference in tag retrieval for the SES categories queried across the three major racial/ethnic groups. Additional studies are required to improve the algorithms' performances and retrieval of semantic tags in multiple populations as well as within different study sites. Indeed, some of the tags developed here (such as "TennCare" in reference to Medicaid) are specific to Tennessee and will require modification to ensure portability regardless of the state in which the algorithms are deployed. Furthermore, these algorithms were created in a de-identified EHR, which required the development of a free text algorithm for insurance status, as the structured insurance information is considered identifying information. An identified EHR may have this insurance information within the structured text. However, the other categories of SES information are likely to only be found within the free text of an identified EHR.

Despite the many challenges faced with the extraction of SES data from the EHR, these algorithms were able to successfully extract a large amount of data not previously accessible for research purposes. The sensitivities, specificities, and PPVs for the algorithms were high considering the limitations of the SES data within the current EHR. Overall, these algorithms represent a first important step in incorporating SES data from EHRs into precision medicine research, as envisioned by the Institute of Medicine and others.

## 5. Resources

Semantic tag and filter lists for each algorithm can be found on the Vanderbilt University Medical Center TREAT Lung Cancer Research Program website (<https://medschool.vanderbilt.edu/treat-lung-cancer-program/>) and the Institute for Computational Biology website ([http://www.icombio.net/?page\\_id=1654](http://www.icombio.net/?page_id=1654)). The code which was used to run the algorithms is available in GitHub.

## 6. Acknowledgements

This work was supported in part by NIH grants U01 HG004798 and its ARRA supplements (DCC) and 1K07CA172294 (MCA). BMH was supported by the NIH/NIGMS Genetics Predoctoral Research Training Program 5T32GM080178-07. The dataset(s) used for the analyses described were obtained from Vanderbilt University Medical Center's BioVU which is supported by institutional funding and by the Vanderbilt CTSA grant funded by the National Center for Research Resources, Grant UL1 RR024975-01, which is now at the National Center for Advancing Translational Sciences, Grant 2 UL1 TR000445-06.

## References

1. *Poverty: Assessing the Distribution of Health Risks by Socioeconomic Position at National and Local Levels.* (2004).
2. T. Seeman *et al.*, *Social Science & Medicine* 66, 72-87 (2008).
3. P. Braveman *et al.*, *Public Health Reports* 129 Suppl 2, 19-31 (2014).
4. National Center for Health Statistics, *Health, United States, 2011: With Special Feature on Socioeconomic Status and Health* (2012).

5. V. Carrieri *et al.*, *Health Econ*, (2016).
6. A. L. Non *et al.*, *American Journal of Public Health* 102, 1559-1565 (2012).
7. M. C. Aldrich *et al.*, *American Journal of Public Health* 103, e73-80 (2013).
8. R. Hauser *et al.*, *Sociological Methodology* 27, 177-298 (1997).
9. K. Fujishiro *et al.*, *Social Science & Medicine* 71, 2100-2107 (2010).
10. in *Kaiser Commission on Medicaid and the Uninsured* T. H. J. K. F. Foundation, Ed. (Washington, D.C., 2012).
11. D. S. Morrison, *International Journal of Epidemiology* 38, 877-883 (2009).
12. National Center for Health Statistics *Health, United States, 2015: With Special Feature on Racial and Ethnic Health Disparities* (2016).
13. T. J. VanderWeele *et al.*, *Epidemiology* 25, 473-484 (2014).
14. S. Cakmak *et al.*, *Journal of Environmental Management* 177, 1-8 (2016).
15. J. S. Kaufman *et al.*, *American Journal of Epidemiology* 181, 464-472 (2015).
16. F. S. Collins *et al.*, *The New England Journal of Medicine* 372, 793-795 (2015).
17. I. S. Kohane, *Nature Reviews. Genetics* 12, 417-428 (2011).
18. J. Basson *et al.*, *American journal of hypertension* 27, 431-444 (2014).
19. IOM (Institute of Medicine), *Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2* (2014).
20. D. M. Roden *et al.*, *Clinical Pharmacology and Therapeutics* 84, 362-369 (2008).
21. J. Pulley *et al.*, *Clinical and Translational Science* 3, 42-48 (2010).
22. D. C. Crawford *et al.*, *Human Heredity* 79, 137-146 (2015).
23. J. B. Hall *et al.*, *PloS one* 9, e99161 (2014).
24. L. Dumitrescu *et al.*, *Genetics in Medicine : official journal of the American College of Medical Genetics* 12, 648-650 (2010).
25. NORC, *Measuring Occupational Prestige on the 2012 General Social Survey* (2014).
26. B. G. Link *et al.*, *J Health Soc Behav Spec No*, 80-94 (1995).
27. T. N. Bethea *et al.*, *Ethnicity & Disease* 26, 157-164 (2016).
28. A. Rawshani *et al.*, *JAMA Internal Medicine*, (2016).
29. "2015 Financial Report " (Vanderbilt University, Nashville, TN. , 2015).