Studying the Confounding Effects of Socio-Ecological Conditions in Retrospective Clinical Research: A Use Case of Social Stress

Matthew K. Breitenstein, PhD[1,2], Jyotishman Pathak, PhD[1], Gyorgy Simon, PhD[1,2]
[1]Mayo Clinic, Division of Biostatistics & Informatics, Rochester, MN; [2]University of Minnesota, Institute for Health Informatics, Minneapolis, MN

Abstract

Socio-ecological Conditions (SECs) are important to include in clinical research models as they have been known to impact the health of patients. However, current clinical research models account for these factors only in an unsatisfyingly rudimentary way. In this study, we developed an SEC Index that captured the latent and direct effects of social stress, one of the many kinds of SEC, on patients' general health as measured by the Charlson Comorbidity Index. We demonstrated that the above SEC Index had a significant effect in a clinical model, a patient-level model with the specific clinical outcome of breast cancer prevalence. Further, we demonstrated that including the SEC Index of social stress into the clinical models significantly increased their performance. Our study demonstrated a viable approach that is interchangeable to include any SEC of interest, to more appropriately account for SECs in clinical research models.

Introduction

Socio-ecological Conditions (SECs) have been known to impact the health of patients, but current clinical research models account for these factors only in an unsatisfyingly rudimentary way. Phenomena such as access to health care and social support networks are examples of SEC; factors that can profoundly impact a patient's health and prognosis once health deteriorates. SECs not only influence outcomes, but they confound patient characteristics and clinical factors, inhibiting the ability of clinical models to estimate the effect of clinical factors independently of SECs. Specifically, without adjusting for SECs we can only estimate the combined effect of SEC and the clinical factors.

In what follows, we describe SECs in more details, present our methodology and demonstrate its utility on a large tertiary care provider in the Midwest U.S. In this study, we set out to develop an SEC Index, a summary of socio-ecological measures that quantifies the effect of SECs on patients' general health. Then we utilize this SEC Index in a clinical risk prediction model with a specific end point (e.g. breast cancer prevalence) and show how the inclusion of the SEC Index results in a statistically significantly better model.

Background

Socio-ecological conditions (SECs) are the embodiment of social and ecological population factors that exist within a defined geographical region (i.e. community) and are known to impact health of an individual patient[1] and enhance stability of retrospective clinical research[2]. A community is an amalgamation of social interactions and geographical proximity exhibiting many confounded and latent characteristics. These characteristics cannot be effectively measured as independent metrics, and are known to vary across geographic regions[1]. While person-level measurements of socioeconomic status are commonly included in statistical analysis in an attempt to control for alternative confounding factors, they do not adequately represent the underlying phenomena at the root of an SEC[4]. Further, placing measures of SEC directly in models has potential to lead to model overfitting and reduction of power. In our study we chose to focus on the SEC phenomena social stress because of its hypothesized relevance to breast cancer and because it has established, validated population measures. However, it is most important to emphasize that any SEC phenomenon of interest, such as factors measuring social contagions, demographic change, and social capital, can be readily substituted within the study design.

Social stress is a phenomenon experienced by individuals when they do not have the resources to address an acute situation[4]. Further, social stress is known to be highly confounded with socioeconomic status (SES)[4]. Social stress has been associated with negative impacts on health, including increased prevalence of asthma, diabetes, gastrointestinal disorders, myocardial infarction, cancer, and rheumatoid arthritis[5]. Social stressors are commonly socially patterned, and can manifest at both the individual and community level[6]. Further, mouse models of social stress have identified correlations between localized (i.e. non-systemic) mammary adipose-specific metabolic changes and increased mammary tumor growth[7].

We chose to use validated measures of social stress: Index of Dissimilarity(D-score)[8], Townsend Index of Socioeconomic Derivation Index(T-score)[9], and poverty[4]. For brevity, we omit the exact definitions, but at a high level, D-score quantifies the homogenization of racial distribution across geographic areas in relation to the entire geographic region of study; and T-score, which is composed of four components, measures unemployment, non-car ownership, household overcrowding, and non-home ownership. Poverty is a measure collected and provided by the US Census Bureau and provides a direct measures of socioeconomic deprivation.

Study Aim

In this work, we seek to understand if variation in established measures of socio-ecological conditions (SEC) for social stress are associated with breast cancer prevalence. Specifically, we developed an SEC Index using the above validated SEC measures of social stress; and later used this index as a covariate in addition to patient-level clinical covariates in a model predicting breast cancer prevalence (clinical model)

Materials

This study utilizes a combination of clinical and population-based data sources. A cohort consisting of primary care patients (n=228,069) with longitudinal clinical data were aggregated from Mayo Clinic's EHR and Enterprise Data Trust[10] using a combination of structured queries. Validated population measures of social stress were calculated using 2010 American Community Survey and US Census datasets. All population data was clustered at the census block group level. A SAS address–census block group crosswalk *(proc geocode)* was used to assign patients to their Census Block Group (CBG) of residence. CBGs (n=278) corresponding to Mayo Clinic's (in Rochester, Minnesota, USA) primary care coverage area (Dodge, Fillmore, Goodhue, Houston, Mower, Olmsted, Wabasha, and Winona Counties in Minnesota, USA) were included in the analysis. Patients who had designated residence in more than one CBG corresponding to Mayo Clinic's primary care coverage area were excluded from this study. Further, to eliminate estimation bias in low coverage areas patients who resided in census block groups with <50 patients were excluded from this study. CBGs were assigned a random identifier and patient-level data was de-identified prior to analysis to ensure patient privacy and confidentiality. The final analysis cohort contained deidentified data for a relatively homogeneous cohort of 94,561 patients and 237 CBGs. A detailed diagram of cohort demographics can be found in **Table 1**.

Table 1: Cohort Demographics (n=94,561)

Age			
	>=18 to ≤25	14,583	15.4%
	>25 to ≤35	16,761	17.7%
	>35 to ≤45	16,262	17.2%
	>45 to ≤55	18,160	19.2%
	>55 to ≤65	11,793	12.5%
	>65 to ≤75	8,393	8.9%
	>75 to ≤85	6,143	6.5%
	>85	2,466	2.6%
Caucasian		85,238	90.1%
Male		43,833	46.4%
Coverage			
	Low (<33%)	17,939	19.0%
	Medium (33 to 50%)	18,893	20.0%
	High (50 to 100%)	57,729	61.0%
Poverty		6,720	7.1%
T-score			
	High Unemployment	15,440	16.3%
	High Non-Car Ownership	6,720	7.1%
D-Score High		9,323	9.9%

Methods

Our proposed method has two steps. First, we develop the SEC Index using a generic endpoint to quantify the effect of the SEC measures on health in general. For this step, we utilize 30% of the data set. In the second step, we utilize the remaining 70% of the data and build the specific clinical model, which includes the SEC Index as an independent variable, with breast cancer prevalence as the clinical end point. Note that to avoid model overfitting, the portion of the data (30%) on which the SEC Index is developed has no overlap with the portion of the data (70%) that the clinical model is constructed on.

SEC Index Construction

We construct an SEC Index to capture the effect of a number of known measures of social stress on the patients' general health. We quantify patients' general health through the Charlson Comorbidity Index[11], with index value in excess of 3 indicating high risk of mortality (poor health). The independent variables include the D-score[8] (averaged over each census block group and dichotomized[12] into high and low at .5), components of the T-score[9], poverty[13], rurality[14], and coverage. The D-score quantifies homogenization of racial distribution across geographic areas in relation to the entire geographic region of study. CBG poverty is defined as absolute poverty thresholds, a measured and defined by the US Census Bureau[13]. In our study, D-score was calculated for individual patients (stratified at 90th percentile) and then averaged within each census block group. T-score was measured by unemployment (90th

percentile), non-car ownership (20th percentile), household overcrowding (10th percentile), and non-home ownership (90th percentile). Poverty was stratified at the 90th percentile. Coverage was defined as the percentage of the population in the CBG who received their *primary care* at Mayo Clinic. Coverage needs to be adjusted for, as patients in certain CBGs only receive specialty care (such as breast cancer treatment) from Mayo Clinic, falsely suggesting that those regions have disproportionally sick people or with disproportionately better access. The independent variables in the SEC Index are not patient-level variables; they are aggregated to CBG-level as they are aimed to capture CBG-level effects. The SEC Index itself is a binomial propensity score model and the index score is the link-space (linear) prediction from this model.

Applying the SEC Index

To show the effectiveness of the SEC Index, we develop two breast cancer prevalence models on the remaining 70% of the cohort. The first model, our baseline, contained patient measures of age and gender and did not utilize the SEC Index; the second model contained age, gender, and our trained SEC Index. Both models were logistic regression models.

Evaluation

We used concordance as the metric of model performance. For a randomly selected pair of patients, with exactly one of the two having breast cancer, concordance is the probability that the predicted risk of breast cancer is higher for the breast cancer patient than for the one without breast cancer. Concordance is also known as C-statistic or Area under the ROC curve (AUC). 100 replications of bootstrap simulation were used to estimate the model performance for the two clinical models.

In bootstrapping, a simulated data set of the same size as the original is created by sampling the patients of the original data set with replacement. As a result of sampling with replacement, some of the original patients are excluded from the simulated data set and others are included multiple times. The patients excluded are referred to as "out-of-bag" patients and are set aside for validation. In each of the 100 replications, the SEC Index model was constructed (on 30% of the simulated data set) and the two clinical models (one with and one without the SEC Index) were developed on the remaining 70% of the simulated data set as described above. The concordance for the two clinical models was calculated on the out-of-bag (validation) patients. After the 100 replications, we had 100 SEC Index models and 100 pairs of clinical models with 100 pairs of concordance values. A paired t-test was utilized to compare the 100 pairs of concordance measures.

Results

We first present the overall results of the bootstrap simulation. Finally, to offer further insight into our methodology, we also present the SEC Index model and the clinical model of a specific replication.

Bootstrap Simulation

Overall, the model for breast cancer prevalence that included the SEC Index in addition clinical characteristics for age and gender in bootstrap replications (n=100) performed strongly significantly ($t=5.8457, p=6.489 \times 10^{-8}$) better than the model without the SEC Index, indicating that the inclusion of the SEC Index is significantly beneficial to the model performance. Further, in 22% of the individual bootstrap replications, the SEC variable was designated as a statistically significant predictor of breast cancer prevalence.

Table 2: Propensity Training Model

Variable	Estimate	Standard Error	Z Test	P-value
Intercept	-1.8026	0.0225	-79.975	< 2e-16
Poverty	0.0943	0.0381	2.478	0.0132
T-score				
V1	-0.1303	0.0284	-4.595	4.32e-06
V2	0.1914	0.0643	2.978	0.0029
Coverage				
33 to 50% (vs. <33%)	-0.1230	0.0305	-4.034	5.48e-05
50 to 100% (vs. <33%)	-0.2765	0.0254	10.883	< 2e-16
D-Score	0.1971	0.0316	6.248	4.16E-10

Specific Example

We randomly chose 1 of the 22 bootstrap models where the trained SEC Index demonstrated a significant impact on breast cancer prevalence measures **(Table 2)**. This model included measures for poverty, coverage, T-Score (unemployment and non-car ownership), and D-Score. Census block groups with high social stress SEC demonstrated a significant (p=0.0168) detrimental (Beta=0.2948) effect **(Table 3)**.

Table 3: Demonstration of SEC				
Variable	**Estimate**	**Standard Error**	**Z Test**	**P-value**
Intercept	-5.7700	0.3562	-16.197	< 2e-16
Age (years), baseline: ≥18 to ≤25				
>25 to ≤35	1.6063	0.2809	5.719	1.07e-08
>35 to ≤45	2.7252	0.2665	10.225	< 2e-16
>45 to ≤55	3.4421	0.2622	13.130	< 2e-16
>55 to ≤65	4.0885	0.2617	15.624	< 2e-16
>65 to ≤75	4.5772	0.2615	17.501	< 2e-16
>75 to ≤85	4.5463	0.2626	17.316	< 2e-16
>85	4.5198	0.2676	16.889	< 2e-16
Male	-2.3522	0.0676	-34.786	< 2e-16
SEC	**0.2948**	0.1233	2.391	**0.0168**

Discussion

We have demonstrated that the use of an SEC Index for social stress significantly increased the performance for prediction of breast cancer prevalence even in our study region located in the Upper Midwest, where differences among CBGs in terms of SEC are relatively modest. We expect the impact of using the SEC Index to amplify when applied to a region where SEC differences among regions are more pronounced. We wish to emphasize that the purpose of this study is neither to recommend the use of specific SEC measures nor to quantify a neighborhood effect, which has been proven theoretically impossible[15]. Rather, our intent is to better identify clinical effect, which our method successfully accomplished.

Strengths and Limitations

Our proposed method offers significant benefits. The most important benefit is that it helps separate the effect of SECs from the effect of clinical variables: without accounting for SEC, we would have only been able to measure the combined effect of the clinical variables and SEC; accounting for SEC helped elucidate the true effect of the clinical variables. SEC measures are so highly correlated with each other that efforts to separate their effect has been deemed fruitless[15]. Including the individual SEC measures into a clinical model would make overfitting inevitable and would limit degrees of freedom, while including the SEC Index contains the collinearity problem of the SEC measures in the SEC Index model. Further, being able to capture the effect of SEC through the patients' generic health (as measured by the Charlson Comorbidity Index) and being able to use it for a specific clinical end point (breast cancer prevalence) enables large-scale generalizability across organizations and coverage areas. The (arguably imperfect) separation between SECs and the clinical variables that the SEC Index affords helps capture the differences in SECs between organizations and coverage areas, leading to more accurate estimates for the clinical effects. Finally, the proposed methodology also allows us to incorporate additional validated or new measures of SEC that may help better separate the impact of SECs and patient characteristics. The SEC measures used in this study are merely a sample of the measures in existence. Alternative measures of interest, for example social contagions, demographic change, and social capital, can be incorporated into the SEC Index in a straightforward way without causing the clinical model to overfit the data.

Future Work

Despite medicine's rigorous pace of advancement, appropriately capturing SEC patterning of disease remains an important topic. With the advent of harnessing social media data and focus on consumer health informatics it is important to consider the lingering issue of how we can quantify the effect of SEC using relatively stable population-based data through validated measures. Advancing our understanding and utilization of SECs is necessary to advance our understanding of the complex, multifactorial causes of cancer[16].

Conclusion

This study demonstrates a viable approach to account for SECs, including social stress, in retrospective clinical research. An important distinction exists between the utilization of SECs to control for confounding effects in retrospective research and utilizations in population health or clinical decision support, critical considerations remains in how to accurately address the long-standing concerns of social epidemiology in consumer health informatics.

Acknowledgements

Data support services provided by Minnesota Population Center. National Historical Geographic Information System: Version 2.0. Minneapolis, MN: University of Minnesota 2011.

References

1. Ed. Kawachi, I., Berkman, L.F., Neighborhoods and Health. Oxford University Press, 2003.
2. Adkins, D.E., Vaisey, S., Toward a Unified Stratification Theory: Structure, Genome, and Status Across Human Societies. Sociological Theory, 2009. 27(2): p.99-121.
3. Oakes, J.M., The measurement of SES in health research: current practice and steps toward a new approach. Social Science & Medicine, 2003. 56(4): p.769-784.
4. Aneshensel, C.S., Social Stress: Theory and Research. Annual Review of Sociology, 1992. 18: p.15-38
5. McEwen, B.S., Stellar, E., Stress and the Individual. Archives of Internal Medicine, 1993. 153: p.2093-2101.
6. DuBois, D.L., Felner, R.D., Brand, S., Adan, A.M., Evans, E.G., A Prospective Study of Life Stress, Social Support, and Adaptation in Early Adolescence. Child Development, 1992. 63(3): p.542-557.
7. Volden PA, Wonder EL, Skor MN. Chronic Social Isolation is Associated with Metabolic Gene Expression Changes Specific to Mammary Adipose Tissue. Cancer Prevention Research, 2013. 6(7):634-45.
8. Friedman, S., Index of dissimilarity. In V. Parrillo (Ed.), *Encyclopedia of social problems.* Thousand Oaks, CA: SAGE Publications, Inc., 2008. p. 489.
9. Townsend, P., Phillimore, P., Beattie, A. Health and Deprivation: Inequality and the North. Croom Helm, 1988.
10. Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. JAMIA. Mar-Apr 2010;17(2):131-135.
11. Charlson, M.E., Pompei, P., Ales, K.L., MacKenzie, C.R., A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of Chronic Disease, 1987. 40(5): p. 373-383.
12. Gilthrope, M.S., The importance of normalisation in the construction of deprivation indices. Journal of Epidemiology and Community Health, 1995. 49: p.S45-S50.
13. US Census Bureau 2009 Poverty Thresholds. shttp://www.census.gov/hhes/www/poverty/data/threshld/index.html
14. 2010 Census Urban and Rural Classification and Urban Area Criteria. https://www.census.gov/geo/reference/ua/urban-rural-2010.html
15. Oakes JM. The (mis)estimation of neighborhood effects: causal inference for a practicable social epidemiology. Social Science & Medicine, 2004. 58: p.1929-1952.
16. Lynch, S.M., Rebbeck, T.R., Bridging the Gap between Biologic, Individual, and Macroenvironmental Factors in Cancer: A Multilevel Approach. Cancer Epidemiology, Biomarkers & Prevention, 2013. 22(4): p.485-495.