Article

Predicting later life health status and mortality using state-level socioeconomic characteristics in early life

Rita Hamad [a,*], David H. Rehkopf [a], Kai Y. Kuan [b], Mark R. Cullen [a]

[a] Stanford University, Department of Medicine, 1070 Arastradero Road, Palo Alto, CA 94304, USA
[b] Stanford University, Department of Statistics, 390 Serra Mall, Stanford, CA 94305, USA

ARTICLE INFO

ABSTRACT

Studies extending across multiple life stages promote an understanding of factors influencing health across the life span. Existing work has largely focused on individual-level rather than area-level early life determinants of health. In this study, we linked multiple data sets to examine whether early life state-level characteristics were predictive of health and mortality decades later. The sample included 143,755 U.S. employees, for whom work life claims and administrative data were linked with early life state-of-residence and mortality. We first created a "state health risk score" (SHRS) and "state mortality risk score" (SMRS) by modeling state-level contextual characteristics with health status and mortality in a randomly selected 30% of the sample (the "training set"). We then examined the association of these scores with objective health status and mortality in later life in the remaining 70% of the sample (the "test set") using multivariate linear and Cox regressions, respectively. The association between the SHRS and adult health status was $\beta=0.14$ (95%CI: 0.084, 0.20), while the hazard ratio for the SMRS was 0.96 (95%CI: 0.93, 1.00). The association between the SHRS and health was not statistically significant in older age groups at a p-level of 0.05, and there was a statistically significantly different association for health status among movers compared to stayers. This study uses a life course perspective and supports the idea of "sensitive periods" in early life that have enduring impacts on health. It adds to the literature examining populations in the U.S. where large linked data sets are infrequently available.

Introduction

Studies that extend across multiple life stages promote an understanding of the factors that influence health across the life span (Braveman & Barclay, 2009). A growing literature has examined not only the individual-level socioeconomic factors in early life that influence health outcomes (Glymour, Avendaño, Haas & Berkman, 2008; Merkin, Karlamangla, Diez Roux, Shrager & Seeman, 2014; Pereira, Li & Power, 2014; Turrell, Lynch, Leite, Raghunathan & Kaplan, 2007), but also the influence of place. Studies in the U.S. have found that a person's state or region of birth is associated with later life development of cancer, dementia, diabetes, heart disease, and other illnesses (Datta, Glymour, Kosheleva & Chen, 2012; Glymour et al., 2013; Greenberg & Schneider, 1998; Patton, Benjamin, Kosheleva, Curtis & Glymour, 2011). Fewer have examined the specific *characteristics* of early life state-of-residence that are predictive of adult health, although one

recent study found small associations of state socioeconomic characteristics with chronic disease during working life (Rehkopf et al., 2015).

Prior work has suggested multiple types of trajectories through which early life factors may influence health and mortality in later life (Ben-Shlomo, Mishra & Kuh, 2014). "Critical period" and "sensitive period" models assume that an exposure in a time window during fetal life or childhood alters an individual's health trajectory early on (Ben-Shlomo & Kuh, 2002). "Accumulation of risk" models suggest that correlated or uncorrelated exposures across the life course interact additively or synergistically to bring about later disease. Meanwhile, "chains of risk" models hypothesize that initial adverse exposures bring about disease in later life because they increase the risk of additional adverse exposures throughout life (Ben-Shlomo et al., 2014).

Adverse exposures have been conceptualized not only in terms of chemical or metabolic risk factors, but also social factors (Halfon, Larson, Lu, Tullis & Russ, 2014). Numerous studies have begun to examine how early and later life socioeconomic status (SES) interact, and systematic reviews have suggested that childhood SES may be as important in determining later life cause-specific mortality and cardiovascular disease as adulthood SES,

* Corresponding author. Tel.: +1 650 497 3208; fax: +1 650 725 6247.
E-mail addresses: rhamad@stanford.edu (R. Hamad),
drehkopf@stanford.edu (D.H. Rehkopf), kykuan@stanford.edu (K.Y. Kuan),
mrcullen@stanford.edu (M.R. Cullen).

depending on the disease process and contextual factors (Galobardes, Lynch & Smith, 2008; Galobardes, Smith & Lynch, 2006). In general, however, a life course perspective is not frequently applied, and researchers have recently called for increased attention to how socioeconomic exposures are "sustained, exacerbated, or attenuated over time" (Corna, 2013).

Moreover, most studies focus on individual-level socioeconomic factors, with less attention to the ways in which contextual factors interact across the life course. For example, area-level socioeconomic factors during childhood may influence educational and economic opportunities or may be associated with poorer housing and environmental conditions (Bartley, Blane & Montgomery, 1997). With regard to macro-level factors that differ across states and countries, differences in social and economic policies may affect how well the safety net buffers vulnerable individuals from adverse conditions (Currie & Rossin-Slater, 2015; Eikemo, Bambra, Judge & Ringdal, 2008).

In this study, we build upon this prior literature by examining how state characteristics in early life predict health status and mortality decades later (Fig. 1). We use composite indices representing socioeconomic characteristics of early state-of-residence as the predictors of interest. We take advantage of multiple large linked data sets among a cohort of U.S. workers, employing in-sample and out-of-sample models to strengthen results. We adjust for potential mediating individual- and area-level factors during adulthood, testing the hypothesis that early life state environment remains important even after controlling for socioeconomic factors during adulthood.

Methods

Data set

The sample included individuals who were ever employed at Alcoa, a large multi-site U.S. manufacturing firm. The data set was constructed by linking administrative, personnel, and medical claims files for employees who worked at least one day since January 1, 1986, which is the earliest that data are available. These data have been made available to researchers through an ongoing collaboration between the firm and the investigators, and they have been described in detail in prior work (Cullen et al., 2006). While this sample is not nationally representative, it is nevertheless demographically and geographically diverse, with extensive available data that enhance the potential to create data linkages across the life span.

Early life state-of-residence was imputed for each individual using the first three digits of his or her Social Security number, a technique commonly implemented based on the fact that these digits differ according to the state in which the Social Security card was issued (Block, Matanoski & Seltser, 1983; Puckett, 2009). Less than 1% were missing information on early life state-of-residence, resulting in a sample size of 143,755 that included individuals from all 50 states and the District of Columbia.

Outcomes

The first outcome examined was an objective measure of health status determined administratively using claims data. This measure was calculated for individuals based on their International Classification of Diseases (ICD) and Current Procedural Terminology (CPT) codes, and health utilization from the prior year. A score of 1 indicates that an individual's health expenditures are likely to fall at the mean in the following year relative to a nationally representative population; each unit increase indicates a one-fold greater value than the mean. This measure was calculated using a proprietary algorithm originally developed as a medical management tool to forecast expenditures and health utilization, based on Diagnostic Cost Group Hierarchical Condition Category (DxCG-HCC) models (Verisk Health, 2015). However, this measure is increasingly being used in epidemiologic and health services research as a marker of objective overall health (Einav, Finkelstein, Kluender & Schrimpf, 2016; Hamad et al., 2015a; Handel, 2011; Modrek & Cullen, 2013; Modrek et al., 2015), and has previously been shown to predict short- and long-term disease outcomes and mortality (Hamad et al., 2015b). In particular, it is valuable for studies such as this one that rely on secondary data sources in which self-reported health is not available. In fact, prior studies have shown that claims-based measures of objective health status are highly correlated with self-rated health, lending credence to their use as a measure of overall health (DeSalvo et al., 2009; Wang et al., 2000). Because of the right-skew of this variable, the natural logarithm was taken, i.e., log(health).

In this sample, all individuals were covered by similar insurance plans with comprehensive benefits, reducing bias due to differences in insurance coverage. For analyses using this outcome, the sample was restricted to those for whom claims data were available in 2004; this was the year in which the largest number of individuals was employed at Alcoa, maximizing the sample size of those for whom the objective health status could be calculated in a given year. Additionally, we restricted the sample to those who were less than 65 years old in 2004, to exclude those individuals for whom we might not have full claims data due to their utilization of Medicare for insurance coverage. The resulting sample size for these analyses was 55,436.

The second outcome was mortality. This was obtained by linking our sample with the Social Security Administration's Death Master File (Social Security Administration Office of Policy, 1998). This included deaths through September 2011, including individuals no longer employed at Alcoa. Remaining individuals were
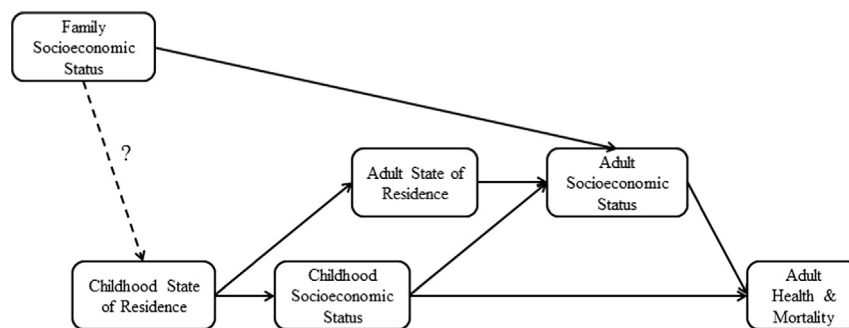
Fig. 1. Conceptual Model: State-of-Residence and Health over the Life Course Note: This figure illustrates hypothesized pathways linking childhood state-of- residence with adult health and mortality. While family socioeconomic status (SES) during childhood has been strongly linked with childhood *neighborhood*- of-residence, it is less likely that it confounds the relationship between childhood *state*-of-residence and adult SES.

censored after this date. The number of deaths was 9678, and these analyses employed the full sample of 143,755 individuals. The average age of death was 66.9 (SD 13.6).

Construction of predictive risk scores

We constructed two composite variables representing the "state health risk score" and "state mortality risk score" of an individual's early life state-of-residence using a predictive algorithm. Conceptually, each of these variables is intended to represent the degree of socioeconomic disadvantage in a state that may lead to worsened health and greater mortality. To construct these, we first collected data on six state characteristics available in U.S. Census data that represent socioeconomic conditions in an individual's early life state-of-residence. These included state unemployment rate, median income in year 2000 dollars, percentage with less than a high school education, percent urban, percent of the population that was white, and Gini coefficient to capture income inequality. As these variables are collected by the Census every 10 years, and because Social Security cards during this period were typically issued at the time of first employment, each individual was linked to the Census variables in the decade in which he or she turned 15 years old. For example, someone born in 1929 was 15 years old in 1944, and would therefore be assigned the Census variables collected in 1940. Each of these variables was standardized with a mean of zero and standard deviation of one, to allow for comparability across measures. The Pearson's coefficients for the pairwise correlations between these variables ranged from 0.02 to 0.87.

In models for each outcome, four out of the six state characteristics were statistically significantly associated with each of the outcomes at a p-level of 0.05 (Supplemental Table 1). Yet the characteristics that were associated with worsened health or greater mortality were not necessarily those that one would expect *a priori*. This is consistent with findings from previous work (Rehkopf et al., 2015), and suggests that these variables may be markers or surrogates of unobserved underlying characteristics of states, such that a causal interpretation should not be applied. Based on this previous work, we therefore chose to construct the composite variables as described above rather than interpreting each characteristic's coefficient individually. Yet developing a composite index in a given data set and then applying this index within the same data set may result in over-fitting of the model to the data; to avoid this, we therefore used a random 30% training subset as described below to construct an *a posteriori* measure of state health or mortality risk for the remaining 70% of the sample. Using very large data sets may also result in over-powered analyses, exacerbating the problem of over-fitting the model to the data (Lenth, 2001). Cross-validation can often address this problem, testing whether a model fit on a random subset of a data set (i.e., the "training set" or "in-sample" group) produces similar results when applied to another subset of the data that was not used in the initial model (i.e., the "test set" or "out-of-sample" group) (Gareth, Witten, Hastie & Tibshirani, 2013). Intuitively, this is akin to performing an additional study in a separate sample to confirm results. Random subsetting of the data set can be accomplished using any statistical programming package, and in this case was done in Stata MP version 14 (College Station, TX).

Consequently, we used these six standardized Census measures as independent variables in a multivariate linear regression model predicting an individual's log(health) among a random 30% of the sample, the training set. This is shown in Eq. (1), in which the variables are indexed by individual (i), state (s), and Census decade of birth (t).

$$\log(\text{Health}_{ist}) = \beta_0 + \beta_1 \text{Unemp}_{st} + \beta_2 \text{Inc}_{st} + \beta_3 \text{Educ}_{st} + \beta_4 \text{Urb}_{st}$$

$$+ \beta_5 \text{White}_{st} + \beta_6 \text{Gini}_{st} + \varepsilon_{ist} \quad (1)$$

We then used the predicted coefficients from this model to predict a "state health risk score" (SHRS) for the remaining 70% of the sample, the test set. The value of this SHRS essentially represents the predicted value of the outcome for individuals in the test set – i.e., $\log(\widehat{\text{Health}})$ – based on the coefficients from the training set model, with higher levels of the index predicting worsened health status. We then repeated this procedure with mortality as the outcome using a Cox proportional hazards model to produce a "state mortality risk score" (SMRS), in which higher values are predictive of greater mortality. This strategy, in which the risk scores are predicted using a subset of the data and then employed in a separate subset, is a form of internal validation of our model. This reduces the chance that our findings are due to over-fitting of the model to our data, although it is not as ideal as external validation on an entirely separate sample (Altman, Vergouwe, Royston & Moons, 2009).

For the six Census measures that were used to construct the SHRS and SMRS, the Cronbach's alpha was 0.80 in the health sample and 0.74 in the mortality sample. This indicates an acceptable level of internal consistency between these variables, suggesting that they are likely capturing the same construct. Similar composite indices have been used in previous studies to capture area-level socioeconomic disadvantage (Krieger et al., 2002; Messer et al., 2006).

Covariates

Individual-level covariates included gender and race (white, black, Hispanic, other). To flexibly adjust for age in the log(health) models, we also controlled for age and age-squared; in the mortality models, we controlled for 10-year birth cohorts.

As described below, we conducted secondary analyses that adjusted for specific adulthood SES measures representing potential mediating pathways (Fig. 1). The first of these was an individual's employment status, dichotomized as hourly or salaried. At Alcoa, hourly workers are more likely to have lower educational attainment than salaried workers. The second adulthood SES measure was an indicator variable representing the individual's state-of-residence during work life, which was determined based on the location at which he or she was employed while at Alcoa.

Data analysis

Primary analyses

The following analyses were conducted in the 70% test subset of the sample. First, we carried out multivariate linear regressions with log(health) as the outcome, controlling for gender, race, age, and age-squared (Eq. (2)). For mortality, we conducted Cox proportional hazards models, controlling for gender, race, and 10-year birth cohorts, with age as the baseline hazard. These are referred to as Model 1 in the accompanying tables.

$$\log(\text{Health}_{ist}) = \alpha_0 + \alpha_1 \text{SHRS}_{st} + \alpha_2 \text{Gender}_i + \alpha_3 \text{Race}_i + \alpha_4 \text{Age}_i$$

$$+ \alpha_5 \text{Age}_i^2 + \varepsilon_{ist} \quad (2)$$

To adjust for possible mediating pathways, we then carried out two additional analyses. In the first, we added employment status to Model 1 above (Model 2). Next, we added indicator variables representing state-of-residence during work life (Model 3). The assignment of both employment status and state-of-residence during work life occurred temporally prior to the assessment of the health outcomes.

In all models, standard errors were clustered by early life state-of-residence and were calculated using the Huber-White sandwich

estimator to be robust to heteroskedasticity. We also carried out a complementary analysis using multi-level models with random intercepts for childhood state-of-residence. For the Cox model, this was implemented with a shared frailty model with gamma distribution (Gutierrez, 2002). The full results from these models are presented in the Supplement and described in the Results section. For multi-level models, we did not carry out Model 3, which included indicator variables for state-of-residence during work life. The large amount of overlap between early and work life states-of-residence and the large number of parameters in these models prevented them from converging.

Sensitivity analyses

We conducted a set of sensitivity analyses comparing those whose early life and work life states-of-residence differed (i.e., "movers," about a quarter of the sample) with those whose states-of-residence did not differ (i.e., "stayers"). To do so, we ran Models 1–3 including an interaction term representing moving status. Outcomes among movers may be expected to differ from stayers due to differences in health status, personality, opportunity, or other factors (Boyle, Halfacree & Robinson, 2014; Larson, Bell & Young, 2004; Sampson & Sharkey, 2008).

To test whether the association between early life state characteristics and overall health became diluted with age, we carried out the analyses for Model 1 after first stratifying the sample by age group (< 46, 46–55, and > 55 years).

Another possible mediating pathway between childhood SES and adult health is an individual's income during adulthood. We were able to observe an individual's wages, obtained from his or her W-2 forms each year, although this was only available for individuals employed during 2002–2012. This resulted in a smaller sample for the analyses of health (N=48,380) and mortality (N=93,646); the number of deaths was 1945, and the mean age of death was 55.8 years. In this restricted sample, we again constructed state health and mortality risk scores. We then conducted Model 1 as described above, and again sequentially added covariates representing wages (Model 2), employment status (Model 3), and work state indicators (Model 4). For analyses examining objective health status, we included wages from 2004. For analyses examining mortality, we took the average of an individual's wages from his or her employment at Alcoa during 2002–2012. Given the right-skew of this variable, the natural logarithm was taken, i.e., log(wages). We again tested whether there was a difference among movers and stayers. For mortality, these analyses should be interpreted with caution given the young age at death and the subsequent non-representative nature of the sample.

Results

Sample characteristics

In both the health and mortality samples, about 20% of the sample was female and three-quarters was white (Table 1). The mean age in 2004 was 45.8 years, and the mean wages were about USD 45,000. About 65% were hourly workers. The mean health score was 1.18, and 6.7% of the sample died.

State health risk score

In Model 1, higher levels of the early life SHRS were associated with worsened health status during work life in the test set (β=0.13, 95%CI: 0.070, 0.19) (Table 2). The magnitude of this association remained largely unchanged when controlling for possible mediating factors including employment status and work life state-

Table 1
Sample characteristics.

Variable	Health Sample N=55,436	Mortality Sample N=143,775
Female (%)	21.8	24.3
Race (%)		
White	79.3	76.8
Black	11.7	12.2
Hispanic	6.6	7.7
Other	2.5	3.2
Age in 2004 (mean ± SD)	45.8 ± 11.2	N/A
Hourly employment status (%)	66.1	62.6
Wages (mean ± SD, in USD)	44,635 ± 60,643	36,158 ± 52,416
Health score in 2004 (mean ± SD)	1.18 ± 1.75	N/A
Died (%)	N/A	6.7

Note: Sample includes employees at Alcoa for whom we have administrative data and information on early life state-of-residence. Wages for health sample are from 2004. Wages for mortality sample are average wages during employee's tenure at Alcoa during 2002–2012. Wages were available for a restricted subset of the health sample (N=48,380) and mortality sample (N=93,646).

Table 2
Association of Early Life State Health Risk Score with 2004 Objective Health Status (N=38,850).

	Beta Coefficient [95% CI]		
	Model 1	Model 2	Model 3
State health risk score	0.13** [0.070, 0.19]	0.13** [0.075, 0.18]	0.12** [0.059, 0.17]
Female	0.35** [0.32, 0.38]	0.36** [0.34, 0.39]	0.36** [0.34, 0.39]
Race (ref=white)			
Black	0.0052 [−0.014, 0.024]	−0.0074 [−0.025, 0.010]	−0.0083 [−0.027, 0.011]
Hispanic	−0.033* [−0.059, −0.0071]	−0.043** [−0.071, −0.015]	−0.042** [−0.065, −0.018]
Other	−0.093** [−0.13, −0.060]	−0.093** [−0.12, −0.062]	−0.087** [−0.12, −0.053]
Age	−0.0076** [−0.012, −0.0029]	−0.0062** [−0.011, −0.0017]	−0.0059** [−0.0096, −0.0023]
Age-squared	0.00046** [0.00041, 0.00051]	0.00045** [0.00040, 0.00050]	0.00045** [0.00041, 0.00050]
Hourly emp status		0.067** [0.048, 0.086]	0.065** [0.049, 0.082]
Constant	−0.91** [−1.03, −0.78]	−0.99** [−1.11, −0.87]	−1.08** [−1.24, −0.93]
Work state indicators	No	No	Yes
R-squared	0.37	0.37	0.38

Sample includes employees at Alcoa for whom administrative and claims data were available in 2004. Analyses were carried out on the 70% test subset of the sample using multivariate linear regression, with robust standard errors clustered by early life state-of-residence. State health risk score was constructed using a 30% training subset of the larger sample using standardized measures of early life state unemployment, median income, percentage with less than a high school education, percent urban, percent white, and Gini coefficient. Health status was calculated from claims data using a third-party algorithm.

* $P < 0.05$.
** $P < 0.01$.

of-residence (Models 2–3). Both of these adult SES factors were also associated with objective health status in each model.

State mortality risk score

In Model 1, the association between early life SMRS and mortality included the null (HR=0.96; 95%CI: 0.93, 1.00) (Table 3). The

Table 3
Association of Early Life State Mortality Risk Score with Mortality (*N*=100,687).

	Hazard Ratio [95% CI]		
	Model 1	Model 2	Model 3
State mortality risk score	0.96*	0.97*	0.97
	[0.93, 1.00]	[0.93, 1.00]	[0.93, 1.00]
Female	0.73**	0.74**	0.76**
	[0.67, 0.80]	[0.68, 0.82]	[0.69, 0.84]
Race (ref=white)			
Black	1.22**	1.16**	1.13**
	[1.12, 1.32]	[1.08, 1.25]	[1.06, 1.21]
Hispanic	0.74**	0.71**	0.68**
	[0.60, 0.91]	[0.58, 0.87]	[0.56, 0.82]
Other	0.73	0.73	0.70*
	[0.52, 1.04]	[0.52, 1.04]	[0.49, 1.00]
Hourly emp status		1.21**	1.20**
		[1.15, 1.27]	[1.14, 1.26]
Work state indicators	No	No	Yes

Sample includes employees at Alcoa for whom we have administrative data. Analyses were carried out on the 70% test subset of the sample using Cox proportional hazards models, with robust standard errors clustered by early life state-of-residence. Additional controls included 10-year birth cohort. State mortality risk score was constructed using a 30% training subset of the larger sample using standardized measures of early life state unemployment, median income, percentage with less than a high school education, percent urban, percent white, and Gini coefficient.

* $P < 0.05$.
** $P < 0.01$.

HR was similar when adjusting for adult employment status (Model 2) and work state indicators (Model 3).

Comparing movers and stayers

In analyses comparing movers and stayers, the association between the SHRS and health was significantly diminished for movers (Model 1) (Table 4). The results were similar when controlling for employment status and work state-of-residence (Models 2–3).

For mortality, the results for movers and stayers were again null (Table 5). This relationship was robust to the inclusion of adult SES variables including employment status and work state-of-residence (Models 2–3).

Sensitivity analyses

When stratifying Model 1 by age group, we found that the SHRS was associated with health status among the youngest age group (β=0.10, 95%CI: 0.043, 0.16), but that this association was less precisely estimated among older individuals (Supplemental Table 2). The R-squared values among older employees were also substantially less: 0.01 and 0.04 compared to 0.24 among the youngest employees.

In analyses including wages in the restricted sample for whom these data were available, the magnitude of the association between the SHRS and log(health) remained similar to that in the primary sample (Model 1) (Supplemental Table 3), even when controlling for wages and other potential mediating variables (Models 2–4). Wages themselves were associated with health in each model. For mortality, there was no association between the state SMRS and mortality in any model (Supplemental Table 4). In this restricted sample, subgroup analyses regressing health status on the SHRS demonstrated diminished effect sizes among movers, similar to what was observed in the larger sample (Supplemental Table 5). For mortality, wide confidence intervals precluded the ability to determine precise estimates, and there were no significant differences between movers and stayers (Supplemental

Table 4
Association of Early Life State Health Risk Score with 2004 Objective Health Status among Movers vs. Stayers (N=38,850).

	Beta Coefficient [95%CI]		
	Model 1	Model 2	Model 3
State health risk score	0.14**	0.14**	0.13**
	[0.076, 0.20]	[0.080, 0.19]	[0.072, 0.19]
Moved	−0.025*	−0.011	−0.011
	[−0.041, −0.0086]	[−0.027, 0.0044]	[−0.026, 0.0031]
State health risk score × Moved	−0.039*	−0.035	−0.036*
	[−0.077, −0.00066]	[−0.072, 0.0022]	[−0.071, −0.00030]
Female	0.35**	0.36**	0.36**
	[0.32, 0.38]	[0.34, 0.39]	[0.34, 0.39]
Race (ref=white)			
Black	0.0053	−0.0072	−0.0083
	[−0.014, 0.024]	[−0.025, 0.010]	[−0.027, 0.011]
Hispanic	−0.033*	−0.043**	−0.042**
	[−0.060, −0.0065]	[−0.071, −0.015]	[−0.066, −0.018]
Other	−0.091**	−0.093**	−0.087**
	[−0.12, −0.060]	[−0.12, −0.062]	[−0.12, −0.053]
Age	−0.0078**	−0.0064**	−0.0064**
	[−0.013, −0.0030]	[−0.011, −0.0019]	[−0.0100, −0.0029]
Age-squared	0.00047**	0.00045**	0.00045**
	[0.00042, 0.00052]	[0.00041, 0.00050]	[0.00041, 0.00050]
Hourly emp status		0.066**	0.064**
		[0.047, 0.085]	[0.048, 0.081]
Constant	−0.90**	−0.98**	−1.07**
	[−1.03, −0.77]	[−1.10, −0.86]	[−1.22, −0.91]
Work state indicators	No	No	Yes
R-squared	0.37	0.37	0.38

Sample includes employees at Alcoa for whom administrative and claims data were available in 2004. Movers are those whose early life and work life states-of-residence differed. Analyses were carried out on the 70% test subset of the sample using multivariate linear regression, with robust standard errors clustered by early life state-of-residence. State health risk score was constructed using a 30% training subset of the larger sample using standardized measures of early life state unemployment, median income, percentage with less than a high school education, percent urban, percent white, and Gini coefficient. Health status was calculated from claims data using a third-party algorithm.

* $P < 0.05$.
** $P < 0.01$.

Table 6). This was robust to adjustment for wages, employment status, and work state.

In analyses involving multi-level modeling, the fixed effect coefficients were not substantially different from the primary models (Supplemental Tables 7 and 8). In both health and mortality analyses, the random parameters from these models indicated that a small amount of the variance in these associations was due to between-state relative to within-state factors.

Discussion

In this study, we demonstrate that early life state contextual factors are associated with worsened overall health status decades later among a sample of working adults in the U.S. Adjusting for potential mediating factors, including wages, employment status, and state-of-residence during adulthood, does not attenuate this relationship. A subgroup analysis suggests that early life state context is not as strongly associated with health status among movers compared to the sample overall. Meanwhile, we find no association between early life SMRS and mortality. Multi-level

Table 5
Association of Early Life State Mortality Risk Score with Mortality among Movers vs. Stayers ($N = 100,687$).

	Hazard Ratio [95% CI]		
	Model 1	**Model 2**	**Model 3**
State mortality risk score	0.99	1.00	0.96
	[0.69, 1.43]	[0.70, 1.44]	[0.65, 1.40]
Moved	0.97	1.00	1.02
	[0.89, 1.05]	[0.93, 1.08]	[0.95, 1.10]
State health risk score × Moved	1.04	1.03	1.08
	[0.84, 1.29]	[0.84, 1.28]	[0.85, 1.37]
Female	0.73**	0.74**	0.76**
	[0.67, 0.79]	[0.68, 0.82]	[0.70, 0.84]
Race (ref=white)			
Black	1.22**	1.16**	1.13**
	[1.13, 1.32]	[1.07, 1.25]	[1.06, 1.21]
Hispanic	0.74**	0.71**	0.68**
	[0.60, 0.90]	[0.58, 0.86]	[0.56, 0.82]
Other	0.73	0.73	0.70
	[0.51, 1.05]	[0.51, 1.05]	[0.49, 1.01]
Hourly emp status		1.21**	1.21**
		[1.15, 1.27]	[1.15, 1.27]
Work state indicators	No	No	Yes

Sample includes employees at Alcoa for whom we have administrative data. Movers are those whose early life and work life states-of-residence differed. Analyses were carried out on the 70% test subset of the sample using Cox proportional hazards models, with robust standard errors clustered by early life state-of-residence. Additional controls included 10-year birth cohort. State mortality risk score was constructed using a 30% training subset of the larger sample using standardized measures of early life state unemployment, median income, percentage with less than a high school education, percent urban, percent white, and Gini coefficient.

*$P < 0.05$.

** $P < 0.01$.

models suggest that a small amount of the variance in these associations is due to between-state relative to within-state factors.

These findings support the life course model of a "sensitive period" during early life, in which factors during childhood influence health in the medium term (Ben-Shlomo et al., 2014). This type of model suggests that effects during this early window are long-lasting, but that the exposure during other time windows (e.g., after moving) can also influence later disease risk. This contrasts with a "critical period" model, in which *only* exposures during early windows are relevant, and later exposures are not influential (Ben-Shlomo & Kuh, 2002). In the context of this study, this implies that both stayers and movers are influenced by their early life state environment during an early window, but that those who move are also affected by later exposures in their new state-of-residence.

This study contributes to the literature on the relative importance of childhood versus adult SES in determining later life health and mortality. The size of the association between childhood contextual factors and health status is small but robust to the inclusion of variables representing adult SES, making our findings consistent with those of prior studies that found that state- or region-of-residence in childhood has a direct effect on adult chronic disease outcomes even after controlling for adult SES (Rehkopf et al., 2015; Nandi, Glymour, Kawachi & VanderWeele, 2012; Patton et al., 2011). Compared to these prior studies, our use of the state health and mortality risk scores represents an innovative way to describe contextual factors during childhood using predictive modeling to construct a composite index and avoid over-fitting the models to the available data. While these models do not imply a causal association, they identify factors that are

consistently related to health and internally validated. Future studies could replicate this work in other samples to provide external validation.

Several possible mechanisms may account for the relationship between early life state-of-residence and later life health. It may be that childhood state-of-residence is mediated by other adult SES factors for which we do not have complete data on the full cohort. For example, early state-of-residence may influence educational and employment opportunities, thus influencing adulthood socioeconomic circumstance, which itself affects adult health (Case, Fertig & Paxson, 2005). Yet some studies have found that low childhood SES is associated with adult health even when individuals are upwardly mobile throughout their lives (Poulton et al., 2002). This might suggest that childhood state-of-residence influences child health, which then goes on to impact adult health. For example, prior work has found that variations in state economic and social policies – such as social welfare and education investments – may lead to differences in child health outcomes (Black, Devereux & Salvanes, 2008; Haider, Jacknowitz & Schoeni, 2003; Hamad & Rehkopf, 2015, 2016), which sets children on a trajectory with respect to their adult health.

Our finding that the association between childhood state-of-residence and adult health is less pronounced among older individuals may represent a dilution of the effect over time, which is also supported by the lack of an association between the SMRS and mortality. Alternately, prior work has found that the association of health with social factors such as education is different among different cohorts and subgroups (Everett et al., 2013), such that cohort effects or differences in the composition of the sample over time may weaken the association between the predicted risk scores and mortality.

We also find that the results for health status are attenuated among movers compared to the sample overall, which may reflect a decreased exposure to early life state environment and therefore a decreased association with later life health. It may also be that those who resided in more disadvantaged environments during childhood were more likely to move, or that they possessed unique personal or family characteristics that buffered them against the effects of state disadvantage. Prior research supports the idea of selective migration (Norman, Boyle & Rees, 2005; Rogerson & Han, 2002), possibly explaining these findings.

The null findings for mortality stand in contrast with prior work that has found variation in mortality rates according to state and county characteristics (Cullen et al., 2012; Kochanek, Murphy & Xu, 2015). These previous studies, however, examined the relationship between mortality and place-of-residence at time of death, while our study employs state-of-residence during childhood. It may be that early life state characteristics have a small effect on mortality many decades later that cannot be detected even in a sample of this size. Yet prior research has found associations between place-of-birth and cause-specific mortality, including death from cardiovascular disease, dementia, atrial fibrillation, and prostate cancer (Datta et al., 2012; Glymour et al., 2013; Glymour, Kosheleva, Wadley, Weiss & Manly, 2011; Greenberg & Schneider, 1998; Schneider, Greenberg & Lu, 1997). It may be that place-of-birth influences the type of illness to which an individual eventually succumbs, but not mortality more generally. For example, prior work has shown that area-of-residence is associated with rates of smoking and cervical cancer screening (Datta et al., 2006a, 2006b). These mediating factors may influence the distribution of causes of death in a given region, but not the overall mortality rates. Finally, we note that the specific state-level socioeconomic factors are associated with mortality in the training set (Supplemental Table 1); a strength of our approach is the process of internal validation, which suggests that narrow confidence intervals do not reflect the stability and replicability of the

estimates even in a similar sample. In other words, we conclude that early life state characteristics are not as predictive of mortality as they are of later life health.

Strengths and limitations

This study has several strengths. First, the data set includes linked individual- and state-level variables among a sample of almost 150,000 individuals, with decades of follow-up across several life stages. Few such data sources are available in the U.S., making this a unique opportunity to contribute to the literature on life course epidemiology. The richness of these data enhances our ability to adjust for the potential mediating pathways between childhood SES and adult health. Also, we use objective measures of health that are less likely to suffer from reporting bias compared to self-reported measures collected in surveys.

There are several limitations to this work. First, although the sample is diverse, it is not nationally representative, and external validity is therefore limited. Second, some researchers may be more interested in the use of more granular geographic regions than state. For example, prior studies have examined the impact on health of residence in a particular county or census tract (Wight et al., 2013; Winkleby & Cubbin, 2003). Even so, states remain a subject of interest for many, particularly as they are considered "laboratories" for policymaking in the U.S. (Brandeis, 1932). There may also be measurement error in our strategy to identify early life state-of-residence using SSNs, in that we are not able to precisely identify the year in which an individual's Social Security card was issued, and how much of early life was spent in this state; this potential misclassification is likely to bias our results to the null. Also, despite the longitudinal nature of the data, this is an observational study, curtailing the ability to make causal inferences. Namely, it may be that early life state-of-residence is endogenous, such that those whose families are more socioeconomically disadvantaged or less healthy are more likely to reside in a more disadvantaged state. While endogeneity is likely to produce more bias when examining neighborhood- or county-of-residence (rather than state), the results should nevertheless be interpreted with caution. Finally, the objective measure of health status we employed was not designed explicitly for the purpose of capturing overall health status. Yet while it was initially created to predict healthcare utilization behavior, it has nevertheless been increasingly employed for purposes similar to our own (Ash et al., 2003; Einav et al., 2016; Petersen, Pietz, Woodard & Byrne, 2005; Shulan, Gao & Moore, 2013).

Conclusions

In this study, we find that early life state contextual factors are associated with health status later in life in the overall sample, but not with mortality. We find that this association is robust to the inclusion of adult measures of SES, suggesting the long-lasting effects of place on health. This study adds to the body of evidence supporting a life course perspective in social epidemiology, and in particular to the idea of "sensitive periods" in early life that have enduring impacts on health. It adds to the literature examining populations in the U.S., where large linked data sets are less frequently available.

Data sharing

As an alternative to providing a de-identified data set to the public domain, we allow access for the purpose of re-analyses or appropriate follow-up analyses by any qualified investigator willing to sign a contractual covenant with the investigators' host institution limiting use of data to a specific agreed upon purpose and observing the same restrictions as are limited in the investigators' contract with Alcoa, such as 60-day manuscript review for compliance purposes.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.ssmph.2016.04.005.

References

Altman, D. G., Vergouwe, Y., Royston, P., & Moons, K. G. M. (2009). Prognosis and prognostic research: validating a prognostic model. *BMJ* (p. 338), 338.

Ash, A. S., Posner, M. A., Speckman, J., Franco, S., Yacht, A. C., & Bramwell, L. (2003). Using claims data to examine mortality trends following hospitalization for heart attack in Medicare. *Health Services Research*, 38, 1253–1262.

Bartley, M., Blane, D., & Montgomery, S. (1997). Socioeconomic determinants of health: health and the life course: why safety nets matter. *BMJ*, 314, 1194.

Ben-Shlomo, Y., & Kuh, D. (2002). A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. *International Journal of Epidemiology*, 31, 285–293.

Ben-Shlomo, Y., Mishra, G., & Kuh, D. (2014). Life course epidemiology In: W. Ahrens, & I. Pigeot (Eds.), *Handbook of epidemiology* (2nd Edition). New York: Springer.

Black, S. E., Devereux, P. J., & Salvanes, K. G. (2008). Staying in the classroom and out of the maternity ward? The effect of compulsory schooling laws on teenage births. *The Economic Journal*, 118, 1025–1054.

Block, G., Matanoski, G., & Seltser, R. (1983). A method for estimating year of birth using social security number. *American Journal of Epidemiology*, 118, 377–395.

Boyle, P., Halfacree, K. H., & Robinson, V. (2014). *Exploring contemporary migration*. Routledge.

Brandeis, J. (1932). *New State Ice Co. v.*, 262. U.S.: Liebmann.

Braveman, P., & Barclay, C. (2009). Health disparities beginning in childhood: a life-course perspective. *Pediatrics*, 124, S163–S175.

Case, A., Fertig, A., & Paxson, C. (2005). The lasting impact of childhood health and circumstance. *Journal of Health Economics*, 24, 365–389.

Corna, L. M. (2013). A life course perspective on socioeconomic inequalities in health: a critical review of conceptual frameworks. *Advances in life course research*, 18, 150–159.

Currie, J., & Rossin-Slater, M. (2015). Early-life origins of life-cycle well-being: research and policy implications. *Journal of Policy Analysis and Management*, 34, 208–242.

Cullen, M. R., Cummins, C., & Fuchs, V. R. (2012). Geographic and racial variation in premature mortality in the U.S.: Analyzing the disparities. *PloS One*, 7, e32930.

Cullen, M. R., Vegso, S., Cantley, L., Galusha, D., Rabinowitz, P., Taiwo, O., et al. (2006). Use of medical insurance claims data for occupational health research. *Journal of Occupational and Environmental Medicine*, 48, 1054–1061.

Datta, G. D., Colditz, G. A., Kawachi, I., Subramanian, S. V., Palmer, J. R., & Rosenberg, L. (2006a). Individual-, neighborhood-, and state-level socioeconomic predictors of cervical carcinoma screening among U.S. black women. *Cancer*, 106, 664–669.

Datta, G. D., Glymour, M. M., Kosheleva, A., & Chen, J. T. (2012). Prostate cancer mortality and birth or adult residence in the southern United States. *Cancer Causes and Control*, 23, 1039–1046.

Datta, G. D., Subramanian, S. V., Colditz, G. A., Kawachi, I., Palmer, J. R., & Rosenberg, L. (2006b). Individual, neighborhood, and state-level predictors of smoking among US Black women: a multilevel analysis. *Social Science and Medicine*, 63, 1034–1044.

DeSalvo, K. B., Jones, T. M., Peabody, J., McDonald, J., Fihn, S., Fan, V., et al. (2009). Health care expenditure prediction with a single item, self-rated health measure. *Medical Care*, 47, 440–447.

Eikemo, T. A., Bambra, C., Judge, K., & Ringdal, K. (2008). Welfare state regimes and differences in self-perceived health in Europe: a multilevel analysis. *Social Science and Medicine*, 66, 2281–2295.

Einav, L., Finkelstein, A., Kluender, R., & Schrimpf, P. (2016). Beyond Statistics: The Economic Content of Risk Scores. Working Paper 21304. *American Economic Journal: Applied Economics,* 8, 195–224.

Everett, B. G., Rehkopf, D. H., & Rogers, R. G. (2013). The nonlinear relationship between education and mortality: an examination of cohort, race/ethnic, and gender differences. *Population research and policy review*, 32, 893–917.

Galobardes, B., Lynch, J. W., & Smith, G. D. (2008). Is the association between childhood socioeconomic circumstances and cause-specific mortality established? Update of a systematic review. *Journal of Epidemiology and Community Health*, 62, 387–390.

Galobardes, B., Smith, G. D., & Lynch, J. W. (2006). Systematic review of the influence of childhood socioeconomic circumstances on risk for cardiovascular disease in adulthood. *Annals of Epidemiology*, 16, 91–104.

Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2013). *5.1 Cross-validation. An Introduction to Statistical Learning*. New York: Springer.

Glymour, M. M., Avendaño, M., Haas, S., & Berkman, L. F. (2008). Lifecourse social conditions and racial disparities in incidence of first stroke. *Annals of Epidemiology*, 18, 904–912.

Glymour, M. M., Benjamin, E. J., Kosheleva, A., Gilsanz, P., Curtis, L. H., & Patton, K. K. (2013). Early life predictors of atrial fibrillation-related mortality: evidence from the health and retirement study. *Health & Place*, 21, 133–139.

Glymour, M. M., Kosheleva, A., Wadley, V. G., Weiss, C., & Manly, J. J. (2011). The geographic distribution of dementia mortality: elevated mortality rates for black and white americans by place of birth. *Alzheimer Disease and Associated Disorders*, 25, 196.

Greenberg, M. R., & Schneider, D. F. (1998). Birth in the South and chronic disease mortality rates in the late twentieth century United States. *Health*, 2, 41–53.

Gutierrez, R. G. (2002). Parametric frailty and shared frailty survival models. *Stata Journal*, 2, 22–44.

Haider, S. J., Jacknowitz, A., & Schoeni, R. F. (2003). Welfare work requirements and child well-being: Evidence from the effects on breast-feeding. *Demography*, 40, 479–497.

Halfon, N., Larson, K., Lu, M., Tullis, E., & Russ, S. (2014). Lifecourse health development: past, present and future. *Maternal and Child Health Journal*, 18, 344–365.

Hamad, R., Modrek, S., Cullen, M.R. (2015a). The Impact of job insecurity on healthcare utilization: Findings from a panel of U.S. Workers. *Health Services Research*, Published online ahead of print, 29 Sept 2015.

Hamad, R., Modrek, S., Kubo, J., Goldstein, B. A., & Cullen, M. R. (2015b). Using "big data" to capture overall health status: Properties and predictive value of a claims-based health risk score. *PloS One*, 10, e0126054.

Hamad, R., & Rehkopf, D. H. (2015). Poverty, pregnancy, and birth outcomes: A study of the Earned Income Tax Credit. *Paediatric and Perinatal Epidemiology*, 29, 444–452.

Hamad, R., Rehkopf, D.H. (2016). Poverty and Child Development: A Longitudinal Study of the Impact of the Earned Income Tax Credit. *American Journal of Epidemiology*, Published online ahead of print, 7 April 2016.

Handel, B.R. (2011). Adverse selection and switching costs in health insurance markets: When nudging hurts. Working Paper 17459. Cambridge, Massachusetts: National Bureau of Economic Research.

Kochanek, K. D., Murphy, S. L., & Xu, J. (2015). Deaths: Final Data for 2011. *National Vital Statistics Reports* (p. 63), 63.

Krieger, N., Chen, J. T., Waterman, P. D., Soobader, M.-J., Subramanian, S. V., & Carson, R. (2002). Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter?: The public health disparities geocoding project. *American Journal of Epidemiology*, 156, 471–482.

Larson, A., Bell, M., & Young, A. F. (2004). Clarifying the relationships between health and residential mobility. *Social Science and Medicine*, 59, 2149–2160.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, 55, 187–193.

Merkin, S. S., Karlamangla, A., Diez Roux, A. V., Shrager, S., & Seeman, T. E. (2014). Life course socioeconomic status and longitudinal accumulation of allostatic load in adulthood: multi-ethnic study of atherosclerosis. *American Journal of Public Health*, 104, e48–e55.

Modrek, S., & Cullen, M. R. (2013). Health consequences of the 'Great Recession' on the employed: evidence from an industrial cohort in aluminum manufacturing. *Social Science and Medicine*, 92, 105–113.

Modrek, S., Hamad, R., & Cullen, M. R. (2015). Psychological Well-Being During the Great Recession: Changes in mental health care utilization in an occupational cohort. *American Journal of Public Health*, 105, 304–310.

Messer, L. C., Laraia, B. A., Kaufman, J. S., Eyster, J., Holzman, C., Culhane, J., et al. (2006). The development of a standardized neighborhood deprivation index. *Journal of Urban Health*, 83, 1041–1062.

Nandi, A., Glymour, M. M., Kawachi, I., & VanderWeele, T. J. (2012). Using marginal structural models to estimate the direct effect of adverse childhood social conditions on onset of heart disease, diabetes, and stroke. *Epidemiology*, 23, 223.

Norman, P., Boyle, P., & Rees, P. (2005). Selective migration, health and deprivation: a longitudinal analysis. *Social Science and Medicine*, 60, 2755–2771.

Patton, K., Benjamin, E. J., Kosheleva, A., Curtis, L. H., & Glymour, M. (2011). Early-life antecedents of atrial fibrillation: place of birth and atrial fibrillation-related mortality. *Annals of Epidemiology*, 21, 732–738.

Pereira, S. M. P., Li, L., & Power, C. (2014). Early-life predictors of leisure-time physical inactivity in midadulthood: findings from a prospective British birth cohort. *American Journal of Epidemiology*, 180, 1098–1108.

Petersen, L. A., Pietz, K., Woodard, L. D., & Byrne, M. (2005). Comparison of the predictive validity of diagnosis-based risk adjusters for clinical outcomes. *Medical Care*, 43, 61–67.

Poulton, R., Caspi, A., Milne, B. J., Thomson, W. M., Taylor, A., Sears, M. R., et al. (2002). Association between children's experience of socioeconomic disadvantage and adult health: a life-course study. *The Lancet*, 360, 1640–1645.

Puckett, C. (2009). The story of the social security number. *Social Security Bulletin*, 69, 55.

Rehkopf, D. H., Eisen, E. A., Modrek, S., MokyrHorner, E., Goldstein, B., & Costello, S. (2015). Early-life state-of-residence characteristics and later life hypertension, diabetes, and ischemic heart disease. *American Journal of Public Health*, 105, 1689–1695.

Rogerson, P. A., & Han, D. (2002). The effects of migration on the detection of geographic differences in disease risk. *Social Science and Medicine*, 55, 1817–1828.

Sampson, R. J., & Sharkey, P. (2008). Neighborhood selection and the social reproduction of concentrated racial inequality. *Demography*, 45, 1–29.

Schneider, D., Greenberg, M. R., & Lu, L. L. (1997). Region of birth and mortality from circulatory diseases among black Americans. *American Journal of Public Health*, 87, 800–804.

Shulan, M., Gao, K., & Moore, C. D. (2013). Predicting 30-day all-cause hospital readmissions. *Health Care Management Science*, 16, 167–175.

Social Security Administration Office of Policy (1998). Announcement of service to epidemiological researchers to provide vital status data on subjects of health research. *Federal Register*, 63, 44942–44944.

Turrell, G., Lynch, J. W., Leite, C., Raghunathan, T., & Kaplan, G. A. (2007). Socioeconomic disadvantage in childhood and across the life course and all-cause mortality and physical function in adulthood: evidence from the Alameda County Study. *Journal of Epidemiology and Community Health*, 61, 723–730.

Verisk Health. (2015). DxCG Risk Analytics. Available at ⟨http://www.veriskhealth.com/answers/population-answers/dxcg-risk-analytics⟩. Accessed 24.08.15.

Wang, M., Rosen, A., Kazis, L., Loveland, S., Anderson, J., & Berlowitz, D. (2000). Correlation of risk adjustment measures based on diagnoses and patient self-reported health status. *Health Services and Outcomes Research Methodology*, 1, 251–265.

Wight, R. G., Aneshensel, C. S., Barrett, C., Ko, M., Chodosh, J., & Karlamangla, A. S. (2013). Urban neighbourhood unemployment history and depressive symptoms over time among late middle age and older adults. *Journal of Epidemiology and Community Health*, 67, 153–158.

Winkleby, M. A., & Cubbin, C. (2003). Influence of individual and neighbourhood socioeconomic status on mortality among black, Mexican-American, and white women and men in the United States. *Journal of Epidemiology and Community Health*, 57, 444–452.