Examining Multi-Level Correlates of Suicide by Merging NVDRS and ACS Data

by

David A Boulifard
Indiana University
Bloomington

Bernice A Pescosolido
Indiana University
Bloomington

CES 17-25 March, 2017

Abstract

This paper describes a novel database and an associated suicide event prediction model that surmount longstanding barriers in suicide risk factor research. The database comingles person-level records from the National Violent Death Reporting System (NVDRS) and the American Community Survey (ACS) to establish a case-control study sample that includes all identified suicide cases, while faithfully reflecting general population sociodemographics, in sixteen USA states during the years 2005-2011. It supports a statistical model of individual suicide risk that accommodates person-level factors and the moderation of these factors by their community rates. Named the *United States Multi-Level Suicide Data Set (US-MSDS),* the database was developed outside the RDC laboratory using publicly available ACS microdata, and reconstructed inside the laboratory using restricted access ACS microdata. Analyses of the latter version yielded findings that largely amplified but also extended those obtained from analyses of the former. This experience shows that the analytic precision achievable using restricted access ACS data can play an important role in conducting social research, although it also indicates that publicly available ACS data have considerable value in conducting preliminary analyses and preparing to use an RDC laboratory. The database development strategy may interest scientists investigating sociodemographic risk factors for other types of low-frequency mortality.

Keyword: suicide risk, multi-level modeling, public use microdata areas, disclosure avoidance

[*] Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau, the CDC, or the individuals acknowledged below. All results have been reviewed to ensure that no confidential information is disclosed. Part of the research reported in this paper was conducted by the first author as a Special Sworn Status researcher at the U.S. Census Bureau's Chicago Research Data Center. Portions of the data were furnished by the National Violent Death Reporting System and the National Center for Health Statistics at the CDC. The authors gratefully acknowledge funding from NIH grant 1R01MH099436; data use permissions granted by the U.S. Census Bureau and the CDC; and invaluable assistance provided by Frank Limehouse at the U.S. Census Bureau, Katie Fowler, Deborah Ingram, and Ilene Rosen at the CDC, Lars Vilhuber at Cornell University, Karen Kafadar at the University of Virginia, and Samuel Harbison and Alex Capshew at Indiana University Bloomington.

1. Introduction

Suicide is recognizable as a major public health problem in the United States, yielding tragic losses of human potential while imposing substantial social and economic burdens. Suicide prevention has become a public health priority (U.S. Department of Health and Human Services, 2012), but the goal of achieving it appears elusive in light of surging rates (Curtin, Warner, and Hedegaard, 2016).

Social scientists maintain that preventive efforts should be informed by understandings of how social environmental factors relate to suicide. Sociologists in particular have drawn inspiration for more than a century from the example of theoretically guided and empirically based research provided by Emile Durkheim's (2006 [1897]) classic study. Durkheim had sought to show that ostensibly individual acts of suicide reflected effects of social contextual phenomena. More specifically, he argued that excesses or deficiencies in social *integration* (participation) and *regulation* (control) accounted for increases in suicide rates.

His ideas continue to figure prominently, albeit with revisions, in suicide theory and research (Wray, Colen, and Pescosolido, 2011) as well as the US government's suicide prevention initiatives (Centers for Disease Control and Prevention, 2008; U.S. Department of Health and Human Services, 2012). This study's principal investigator, Bernice Pescosolido, has reconceptualized and expanded Durkheim's explanation within the context of social network theory, placing the act of suicide at the nexus of interacting individual and social phenomena (Pescosolido, 1994, 2011).

In doing so, it highlights a bifurcation in methods that has hampered modern etiological research. Studies have typically examined individual-level risk factors by comparing their occurrence rates among suicide cases and matched controls, or community-level risk factors by using them to predict observed suicide rates. These approaches incur substantial liabilities when pursued separately: the former neglects potentially significant environmental influences; the latter courts an inferential hazard widely known as the *ecological fallacy* (Robinson, 1950).

The research described here seeks to surmount these limitations by conducting empirical investigations using data that support conjoint evaluations of individual- and community-level risk factors. Obtaining such data, however, presents a challenge. Given the low frequency and wide geographic dispersion of suicide events, acquiring sufficiently many cases to support rich hypothesis testing via cross-sectional or prospective research design is prohibitively expensive. Retrospective design can solve this problem through proactive case accumulation, but it nonetheless requires a means of finding the cases and suitably matched controls.

Responding to this challenge, Pescosolido (2012) proposed the development of a multi-level suicide study database by drawing information on cases, controls, and community attributes from distinct sources. Crucial to the viability of her approach was the availability in two existing federal databases of information suitable for creating the subject data records.

One of these databases is the CDC's *National Violent Death Reporting System* (*NVDRS*), which extensively documents nearly all suicides occurring in about a third of USA states. Rich in risk factor information, these data can directly support descriptive research on suicide cases, but their use for etiological analysis is limited by the absence of comparable information for members of the general population.

The other database is the U.S. Census Bureau's *American Community Survey* (*ACS*), which gathers information formerly obtained via the decennial census long form. Rich in demographic and socioeconomic information, these data can furnish general population controls to complement suicide cases.

With funding from NIH grant 1R01MH099436, a research team directed by Pescosolido at Indiana University did assemble a study database from these and other resources to conduct multi-level risk factor analyses. Specifically, we comingled NVDRS restricted-access records with ACS *Public Use Microdata Sample (PUMS)* records. In doing so, we faced a challenge concerning our use of residence location to define subject communities and their corresponding community-level variables.

A primary feature of PUMS data set design is the adoption of *Public Use Microdata Areas (PUMAs)* as geographic units for residence location. Created as "combinations of contiguous counties or census tracts" having populations of at least 100,000 (U.S. Census Bureau, 2009), PUMAs protect respondent identity because their significant size limits the discriminating power of attribute combinations (Lauger, Wisniewski, and McKenna, 2014).

As units of intrastate geographic subdivision, PUMAs are partly incompatible with counties, which are used for residence location within NVDRS (and many other) data sets. Although they sometimes coincide, either type can subdivide the other. To develop units that could serve in common between the differently sourced records, we subsumed intersecting PUMAs and counties into larger areas, which we called *PUMA Groups*, within which any instances of fragmentation were contained.

Although we obtained significant findings from several hypothesis tests, we wondered if our means of geographic unit reconciliation had sacrificed analytic precision (and with it substantive findings), as the averaging of statistics over combined geographic areas might obscure meaningful distinctions among them. To explore this issue, we sought permission from the Census Bureau to reconstruct our database and repeat our analyses within an RDC laboratory, using ACS data that provided county-level residence location.

In this paper's remaining sections, we describe our efforts inside and outside the RDC laboratory, explaining the design and construction of our study databases, a selection of the statistical hypotheses we tested, and the differences in findings we obtained using public- and restricted-access ACS data sets. We then seek to interpret these differences and assess their implications for conducting further health-related research in a similar vein.

2. Database Development

Figure 1 provides a schematic representation of the database we assembled, which we named the *United States Multi-Level Suicide Data Set (US-MSDS)*. In what follows we discuss the data acquisition and preparation, custom geographic unit formation, and file construction procedure.

		Individual Level					Community Level					
		NVDRS (Case) & ACS (Ctrl)					ACS		USA Counties		RCMS	
Commu-nity	Year	Outcome Status	Obs Weight	Sex	Age	…	Pct Widowed	…	Persons /Sq Mile	…	Prot Cng /10K Pop	…
x	x	Case	x	x	x	…	x	…	x	…	x	…
			…	…	…	…						
		Ctrl	x	x	x	…						
			…	…	…	…						
	x	Case	x	x	x	…	x	…				
			…	…	…	…						
		Ctrl	x	x	x	…						
			…	…	…	…						
…	…	…	…	…	…	…	…	…	…	…	…	…

Figure 1. Schematic representation of analysis database, formed by comingling individual-level (case and control) records and appending community-level variables. Within a given community, each community variable has one value across records for a given year (ACS) or the study period (USA Counties, RCMS).

2.1 Data Acquisition and Preparation

Suicide case records were created from person-level data provided by the *National Violent Death Reporting System* (*NVDRS*), a project administered by the CDC's National Center for Injury Prevention and Control that gathers demographic and incident-related information on violent deaths of participating states' residents. A custom compilation of restricted-access data was obtained for the 16 states having fully operational investigation systems during the study's entire time period, namely the years 2005-2011. These states were: Alaska, Colorado, Georgia, Kentucky, Maryland, Massachusetts, New Jersey, New Mexico, North Carolina, Oklahoma, Oregon, Rhode Island, South Carolina, Utah, Virginia, and Wisconsin.

General population control records were created from person-level data provided by the *American Community Survey* (*ACS*), a project administered by the U.S. Census Bureau that gathers, on a rolling basis in every county,[1] demographic and socioeconomic information on members of randomly selected households. Annual compilations were obtained for the 16 states and seven years noted above. *Public Use Microdata Sample* (*PUMS*) files were downloaded from the Bureau's website for the in-house version; restricted access files were made available to the project for the RDC laboratory version.

[1] We refer as *counties* to geographic areas legally identified by other terms (e.g., parishes, boroughs) that the U.S. Census Bureau considers the "statistical equivalents" of counties.

The NVDRS person file records selected were for residents of the study states who were at least 15 years old and classified by an NVDRS team member (the *abstractor*) as suicides. The ACS person file records selected were likewise for residents of these states who were at least 15 years old. After listwise deletion of missing data, 63,190 (94%) of the NVDRS records selected remained, whereas 4,372,335 (96%) of the public-access and (approximately) 6,510,000 (96%) of the restricted-access ACS records selected remained.[2]

To identify jointly available and compatibly codable risk factor variables for use in these records, we examined variable definitions and descriptive statistics for the NVDRS and ACS source data sets. This effort yielded six factors, namely: sex (male, female), age (in years, grouped as 15-24, 25-44, 45-64, 65-up), race (White, Black, American Indian or Alaska Native, Asian or Pacific Islander), ethnicity (Hispanic, non-Hispanic), national origin (born in USA, not born in USA), and marital status (married, widowed, divorced, separated, never married).

Observation weights, designed to inflate sample statistics to population values, allowed us to generate community-level variables from ACS person-level data. We estimated the percentages of individuals in the community falling within the categories of each risk factor, and the percentage falling below the poverty line.

A second source of community-level variables was the U.S. Census Bureau's publically available *USA Counties (USAC)* database, which provides county-level demographic and socioeconomic statistics such as those previously published in the Bureau's (2007) *County and City Data Book*. We obtained information needed to calculate the percent cumulative 5-year migration as of 2009 and the population density as of 2010 for the community.

A third source was the Association of Statisticians of American Religious Bodies' publically available *Religious Congregations and Membership Study* (*RCMS*) for the year 2010, which provides statistics on 236 religious organizations. These organizations were classified into Steensland's (2000) religious tradition categories, and county-level counts of their congregations were summed accordingly to calculate concentration levels in the community. The categories were: Evangelical Protestant and Mormon, Mainstream Protestant, Black Protestant, Catholic and Orthodox, Jewish, and Other.

A fourth source was Hanzlick's (2007) *Death Investigation: Systems and Procedures*, which provided information to distinguish NVDRS states employing coroners or a mixture of coroners and medical examiners for death investigation (Colorado, Georgia, Kentucky, South Carolina, and Wisconsin) from those employing medical examiners exclusively (the remainder).

A final source of community-level information, obtained for all study years, was the CDC's (National Center for Health Statistics, 2015) *Compressed Mortality File* (*CMF*). Its restricted-access component provides annual, county-level counts of US resident deaths by cause and demographic group; a companion public access component likewise provides population counts. We used this information for observation weight adjustments (described below) that preserved community suicide rates amid the listwise deletions of missing data.

[2] The Census Bureau imputes ACS data where variables are defined but values are missing. Our sample losses resulted from excluding records for which ACS codes indicated membership in more than one of our race groups.

2.2 Geographic Unit Formation

In what follows we describe our effort, needed for the in-house database version, to reconcile partly incompatible geographic units by developing superordinate units called *PUMA Groups*. We also describe an effort, needed for the RDC laboratory version, to accommodate temporal changes in county geography by likewise developing superordinate units we called *County Clusters*.

Figure 2 schematically depicts a hypothetical configuration of PUMAs and counties, using letters to label the areas of their intersection. Instances in which either type of unit subdivides the other create ambiguity in location matching. The blocks of adjacent cells denoted by shading (i.e., AB, CDE, FGHI), show how conjoint clustering of PUMAs and counties can establish units affording unequivocal correspondence, albeit at a loss of geographic specificity.

	County					
PUMA	1	2	3	4	5	6
100	A	B				
201			C			
202			D			
203			E			
300				F	G	
400					H	I

Figure 2. Hypothetical configuration of PUMAs and counties. Letters label the areas of intersection. Shaded blocks of adjacent cells denote PUMA Groups.

To perform this clustering, we obtained information on the relationships between PUMAs and counties, as they were defined in the year 2000, through the Missouri Census Data Center's *MABLE/Geocorr2K* website application. It generated a file with one record for each PUMA-county intersection that listed the area's population count. This enumeration enabled us to frame PUMA-county relationships in statistical rather than topological terms, treating PUMA- and county-based subdivisions as dimensions of a contingency table into which population units could be cross-classified.[3] Figure 2 exemplifies such a table if the letters are taken to be the non-zero population counts.

The clustering process was guided by an effort to minimize the size of PUMA Groups while ensuring that PUMAs and counties were fully nested within them. It was facilitated by the fact that many PUMAs were simply groups or subdivisions of counties, as exemplified in Figure 2 respectively by PUMA 100 and PUMAs 201 through 203. More complex configurations, as exemplified in Figure 2 by PUMAs 300 and 400, required special attention, and sometimes yielded PUMA Groups that were comparatively large.[4]

[3] PUMAs lie wholly within, and collectively exhaust, each state's geographic area. The same is true of counties.

[4] As a practical matter, we used a spreadsheet application to perform the PUMA Group assignments. Automatic formatting techniques helped us to identify and resolve instances of incomplete nesting.

Our effort at geographic unit reconciliation faced an additional challenge in the question of its temporal stability. During the study period, year 2000 PUMA definitions remained in use by ACS PUMS data sets, but changes occurred in county geography (U.S. Census Bureau, 2010). Some of them comprised boundary alterations, but others comprised annexation of some counties by others, or the collective replacement of several counties within an area by a set of incompatibly defined ones.

To assess the significance of these changes, we obtained information on the relationships between year 2000 PUMAs and year 2010 counties from the Missouri Census Data Center's *MABLE/Geocorr12* website application. It showed that all the changes occurred inside the initially defined boundaries of PUMA Groups, and that modest conjoint clustering of year 2000 and year 2010 counties was sufficient to yield County Clusters that would serve as location units in common between county-level data records of all vintages.

Using PUMA Groups as location units imposed substantial consolidation, reducing the number of communities to 289 from the original 534 PUMAs and 963 counties. They ranged in year-2010 population size from a minimum of 97,265 (in New Jersey, where PUMAs invariably aggregated counties) to a maximum of 4,839,852 (in Massachusetts, where PUMAs extensively fragmented counties).

By contrast, using County Clusters as location units imposed only modest consolidation, reducing the number of communities to 957 from the original 963 counties. They ranged in year-2010 population size from a minimum of 558 (for Yakutat borough in Alaska) to a maximum of 1,503,085 (for Middlesex county in Massachusetts).

2.3 File Construction

The five steps listed below summarize the assembly of our multi-level database.

1. Map source file residence location units into their superordinate counterparts, converting PUMAs and counties to PUMA Groups for the in-house version and counties to County Clusters for the RDC laboratory version.

2. Draw case and control records respectively from NVDRS and ACS person-level files, and recode risk factor variables on common bases.

3. Generate community-level records: (a) summarize records from the ACS person file to the level of community and year, and records from the USAC and RCMS files to the level of community alone; (b) compute required ratios within each summarized record.

4. Comingle case and control records, sort them by community and year, and join them with community-level records matched by community and year or community alone.

5. Initialize observation weights for the case and control records respectively as unity and the ACS person weight, and adjust them to match their sums respectively with CMF-based counts of suicide cases and non-cases (population minus cases) by community and year.

3. Data Analysis

A fourfold table in which sampling units are cross-classified on dichotomous antecedent and outcome variables (Figure 3) offers a useful starting point for discussing the study's analytic strategy. Fleiss (1981) discusses designs that can furnish data for such a table, and the types of analyses they can support.

		Outcome		
		Y	N	
Antecedent	Y	a	b	a + b
	N	c	d	c + d
		a + c	b + d	

Figure 3. Fourfold table for dichotomous antecedent and outcome variables.

Cross-sectional design draws a sample from the population overall, whereas prospective design draws one separately from each antecedent subpopulation, and retrospective design draws one separately from each outcome subpopulation. Cross-sectional design permits estimation of the outcome probability conditional upon the antecedent status, or the antecedent probability conditional upon the outcome status, because the sample preserves population proportions along each table dimension. By contrast, prospective design permits only the former type of estimation and retrospective design only the latter.

Our study furnished data that could fill such a table, where the outcome is suicide occurrence during the observation period and the antecedent is any given demographic attribute. Although the data collection method was essentially retrospective, the sample was functionally cross-sectional: each cell count estimated a *population frequency*.[5] We could therefore estimate suicide risk associated with the attribute, and furthermore stratify the sample by community attribute rate to ascertain how that risk varies accordingly. This type of multi-level analysis has value in testing predictions from Pescosolido's reformulation of Durkheim's theory.

Previous research has supported the notion that attributes or circumstances normally associated with personal hardship, such as being widowed, increase suicide risk. The social network perspective implies that these effects are likely to vary with individuals' experiences of social connectedness and support, which in turn are likely to vary with communal norms. Thus, for instance, being widowed may be less difficult to tolerate in the presence of others who share that status.

[5] Treating ACS records solely as controls did overlook the possibility that some surveyed individuals committed suicide after their data had been collected, but the low annual rate of suicide events (about 15 per 100,000) implies that their inclusion had negligible effects on control counts. Moreover, the observation weight adjustments made in constructing the database ensured consistency between column marginals and CMF-reported suicide rates.

In practice we tested this prediction by performing a simultaneous multiple logistic regression of suicide occurrence on several individual-level risk factor variables, their community-level counterparts, and their respective interactions. For example, one set of predictors comprised the individual marital status of being widowed, the community rate of this status, and the product of the individual status and its community rate. Additional individual- and community-level predictors respectively included sex−by−age groups and several environmental factors.

Stated more formally, we fitted a model in which the logarithm of the odds (log-odds) of a suicide event equals a weighted sum of the predictor variables:

$$y = \log(p/(1-p)) = \beta_0 + \sum\beta_i X_i + \varepsilon,$$

where p is the probability of an event during the observation period, each X_i ($i = 1,\ldots,k$) is a predictor, and ε denotes effects unexplained by the predictors.[6] Regarding individual-level effects as fixed and community-level effects as random, we estimated standard errors for predictor coefficients using the Jackknife procedure with communities as clusters, and tested predictor coefficient significance using a false discovery rate of .05.[7]

4. Results

Descriptive statistics for selected predictor variables are presented in Table 1. These values, and descriptive statistics otherwise used in this report, were drawn solely from the in-house database in consequence of RDC laboratory limits on information disclosure. We believe, however, that this limitation only minimally affected our efforts to compare analytic results across database versions, as explained below in our discussion of effects plot construction.

Parameter estimates from a model fitted on selected variables using each database version are presented in Table 2. Coefficients from the two versions were almost perfectly coincident in sign, and (apart from a prominent exception for marital status) generally comparable in magnitude. Standard errors from the laboratory version were smaller in all cases, and on average by 29%.

Marks in the rightmost table column indicate changes in coefficient statistical significance between the two versions. All coefficients having significance in the in-house version retained it in the laboratory version, whereas seven coefficients gained significance in the laboratory version. Of special interest among the latter are coefficients of interaction effects within the predictor sets for American Indian/Alaska Native (AIAN) race and separated marital status, which indicate theoretically relevant moderations of attribute-associated individual risk by community attribute rate that were missed by the in-house analysis.

[6] Conversely, in accordance with this model, the odds of a suicide event are e^y, and the probability is $e^y/(1+e^y)$. The generally low annual rate of suicide makes the odds an excellent approximation to the probability.

[7] All models were fitted using the SAS® Surveylogistic procedure.

Table 1. Descriptive statistics for predictor variables (2005-2011 US-MSDS)

Individual-Level	Min	Max	Mean	Std Dev
Male				
Aged 15-24	0	1	0.09	0.28
Aged 25-44	0	1	0.17	0.37
Aged 45-64	0	1	0.16	0.37
Aged 65 Up	0	1	0.07	0.25
Female				
Aged 15-24	0	1	0.08	0.28
Aged 25-44	0	1	0.17	0.38
Aged 45-64	0	1	0.17	0.38
Aged 65 Up	0	1	0.09	0.29
White	0	1	0.80	0.40
African American	0	1	0.15	0.36
American Indian/Alaska Native	0	1	0.01	0.11
Asian/Pacific Islander	0	1	0.04	0.19
Hispanic	0	1	0.06	0.23
Born in U.S.A.	0	1	0.89	0.32
Married	0	1	0.51	0.50
Widowed	0	1	0.06	0.24
Divorced	0	1	0.11	0.31
Separated	0	1	0.02	0.15
Never Married	0	1	0.30	0.46
Community-Level	Min	Max	Mean	Std Dev
% White	20.42	98.91	74.17	16.06
% African American	0.00	66.36	15.10	15.20
% American Indian/Alaska Native	0.00	68.33	1.15	3.99
% Asian/Pacific Islander	0.00	21.88	3.61	3.69
% Hispanic	0.00	73.01	9.95	9.26
% Born in U.S.A.	53.94	99.78	88.76	8.42
% Married	24.01	68.12	50.59	6.42
% Widowed	1.87	11.73	6.02	1.50
% Divorced	5.10	18.51	10.46	1.93
% Separated	0.24	6.69	2.32	0.96
% Never Married	16.62	55.77	30.61	5.93
% Below Poverty Line	1.70	37.29	13.16	5.45
% Net Migration (5-Yr Cumulative)	-15.79	35.11	4.46	7.68
Persons per Square Mile (Log-10)	-0.63	4.14	2.59	0.70
Medical Examiner System	0	1	0.64	0.48
Congregations/10K Persons				
Evangelical Protestant/Mormon	1.44	28.49	7.46	5.29
Mainline Protestant	0.12	13.62	2.77	1.98
Black Protestant	0.00	7.41	0.69	0.98
Catholic and Orthodox	0.03	13.20	0.70	0.83
Jewish	0.00	1.61	0.11	0.18

Statistics in this table were computed using the in-house version of the merged database, in which communities were defined by Puma Groups. Individual-level attributes were represented using dichotomous predictors in which 1 denoted presence and 0 absence. Regression analysis reference categories for the sex-by-age, race, and marital status attributes respectively were *Male Aged 15-24*, *White*, and *Married*.

Table 2. Comparison of logistic regression model parameter estimates (2005-2011 US-MSDS)

		Puma Group Version		County Cluster Version		Dif
Predictor	Level	Beta	Std Err	Beta	Std Err	Sig
Male						
Aged 25-44	Indv	0.5660*	0.0349	0.5670*	0.0257	
Aged 45-64	Indv	0.6455*	0.0381	0.6453*	0.0284	
Aged 65 Up	Indv	0.6849*	0.0353	0.6828*	0.0299	
Female						
Aged 15-24	Indv	-1.4907*	0.0458	-1.4927*	0.0416	
Aged 25-44	Indv	-0.6893*	0.0370	-0.6832*	0.0305	
Aged 45-64	Indv	-0.5660*	0.0428	-0.5679*	0.0330	
Aged 65 Up	Indv	-1.5632*	0.0523	-1.5585*	0.0433	
African American	Indv	-0.7959*	0.0560	-0.7823*	0.0540	
	Com	0.0043*	0.0016	0.0038*	0.0010	
	I x C	-0.0065*	0.0020	-0.0067*	0.0019	
Amer Indian/Alaska Native	Indv	-0.2287	0.1381	-0.3689*	0.1021	+
	Com	0.0021	0.0038	0.0016	0.0032	
	I x C	0.0135	0.0151	0.0168*	0.0070	+
Asian/Pacific Islander	Indv	-0.5889*	0.0835	-0.5671*	0.0721	
	Com	0.0215*	0.0074	0.0184*	0.0059	
	I x C	0.0235	0.0113	0.0203	0.0099	
Hispanic	Indv	-0.2871*	0.0812	-0.3810*	0.0684	
	Com	0.0121*	0.0035	0.0131*	0.0017	
	I x C	0.0008	0.0036	0.0018	0.0031	
Born in USA	Indv	0.4357	0.4357	0.6515	0.3935	
	Com	0.0253*	0.0064	0.0287*	0.0050	
	I x C	0.0000	0.0050	-0.0029	0.0045	
Widowed	Indv	1.4212*	0.0873	1.5184*	0.0724	
	Com	-0.0371*	0.0160	-0.0226*	0.0060	
	I x C	-0.0730*	0.0149	-0.0884*	0.0116	
Divorced	Indv	1.6607*	0.0837	1.7891*	0.0597	
	Com	0.0395*	0.0068	0.0320*	0.0032	
	I x C	-0.0484*	0.0079	-0.0597*	0.0054	
Separated	Indv	0.3238	0.2847	0.6820*	0.2180	+
	Com	0.0184	0.0133	0.0166*	0.0067	+
	I x C	-0.1349	0.0837	-0.2600*	0.0580	+
Never Married	Indv	0.6221*	0.0928	0.8077*	0.0706	
	Com	-0.0065	0.0032	-0.0060*	0.0019	+
	I x C	0.0022	0.0032	-0.0039	0.0023	
% Below Poverty Line	Com	0.0003	0.0028	0.0014	0.0018	
% Net Migration (5-Yr Cum)	Com	0.0014	0.0020	0.0005	0.0013	
Persons/Sq Mile (Log-10)	Com	-0.0646	0.0351	-0.0371	0.0289	
Medical Examiner System	Com	-0.0153	0.0413	-0.0199	0.0263	
Congregations/10K Persons						
Evangelical Prot/Mormon	Com	0.0136*	0.0049	0.0102*	0.0029	
Mainline Protestant	Com	-0.0010	0.0095	-0.0028	0.0041	
Black Protestant	Com	-0.0254	0.0141	-0.0195*	0.0083	+
Catholic and Orthodox	Com	-0.0083	0.0473	-0.0075	0.0099	
Jewish	Com	-0.0750	0.1104	-0.1372	0.0745	

N cases = 63190, N controls = 4372335/6510000, LR χ^2 = 53279/52956, df = 43, $p < .0001$.
I x C = Individual-Community product term. *$p < .05$ (FDR threshold = .0267/.0349)

Effects plots for these two predictor sets, and the widowed marital status set mentioned above by way of example, are presented in Figure 4. In each case, the community attribute rate is plotted on the horizontal axis and the log-odds of suicide occurrence on the vertical axis. Straight lines show the variation of suicide risk with community attribute rate for individuals belonging respectively to the selected attribute and reference groups. Curved lines denote confidence bands based on a false discovery rate of .05 for the fitted model.

Values for the community attribute rate variable were permitted to range between the sample minimum and maximum, whereas those for variables unconnected with the predictor set were fixed at sample means. We judged the use of in-house database statistics to be acceptable for these purposes in plots for both analysis versions, after using the USAC and CMF data sets to ascertain that geographic unit aggregation only modestly affects full sample mean values of community rate and interaction effect variables.[8]

Within each plot, the slope of the reference group's logit line corresponds to the predictor set's community rate variable coefficient. The difference in slopes between the attribute and reference groups' lines corresponds to the interaction term coefficient. Moderation of the risk difference between groups accordingly manifests in changing vertical distance between the logit lines as horizontal axis values vary.

A statistically significant, negative interaction effect for widowed marital status found in both analysis versions accords with the previously noted expectation. Elevation in suicide risk associated with widowhood clearly decreases as the community widowhood rate increases, although non-overlapping confidence bands at the plot's right-hand margin indicate that the risk remains greater for widowed than married individuals at the 12% maximum [Puma Group] rate. Visibly narrower confidence bands in the laboratory version accord with previously noted decreases in standard errors.

Narrowing of confidence bands is likewise evident in the laboratory version of plots for the separated marital status and the AIAN race predictor sets, along with sharper downward slope for the separated status logit line. The marital status logit lines and their respective confidence bands cross completely, with non-overlapping confidence bands at the left and right margins respectively implying that marital separation can qualify as a risk or protective factor. The race logit lines likewise cross, although confidence bands at the right margin continue to overlap.

These findings fulfilled our expectation of greater precision in the laboratory-based analysis, but we wondered if the improvement was attributable to greater geographic specificity in community definition. The cause, for example, might simply be the greater number of ACS data records available in the laboratory setting. Reviewing our in-house resources, we realized that the CMF data set afforded an opportunity to examine directly how geographic specificity could affect our findings for the AIAN predictor group.

[8] This finding may seem surprising, but a plausible explanation for community rate variables is that their weighted sums yield, in cases such as the marital statuses, the numerators used to compute them, and in cases such as the race groups, a proportion of those numerators that varies modestly across communities. Similar reasoning may likewise explain the less striking but nonetheless substantial stability found for the means of interaction terms.

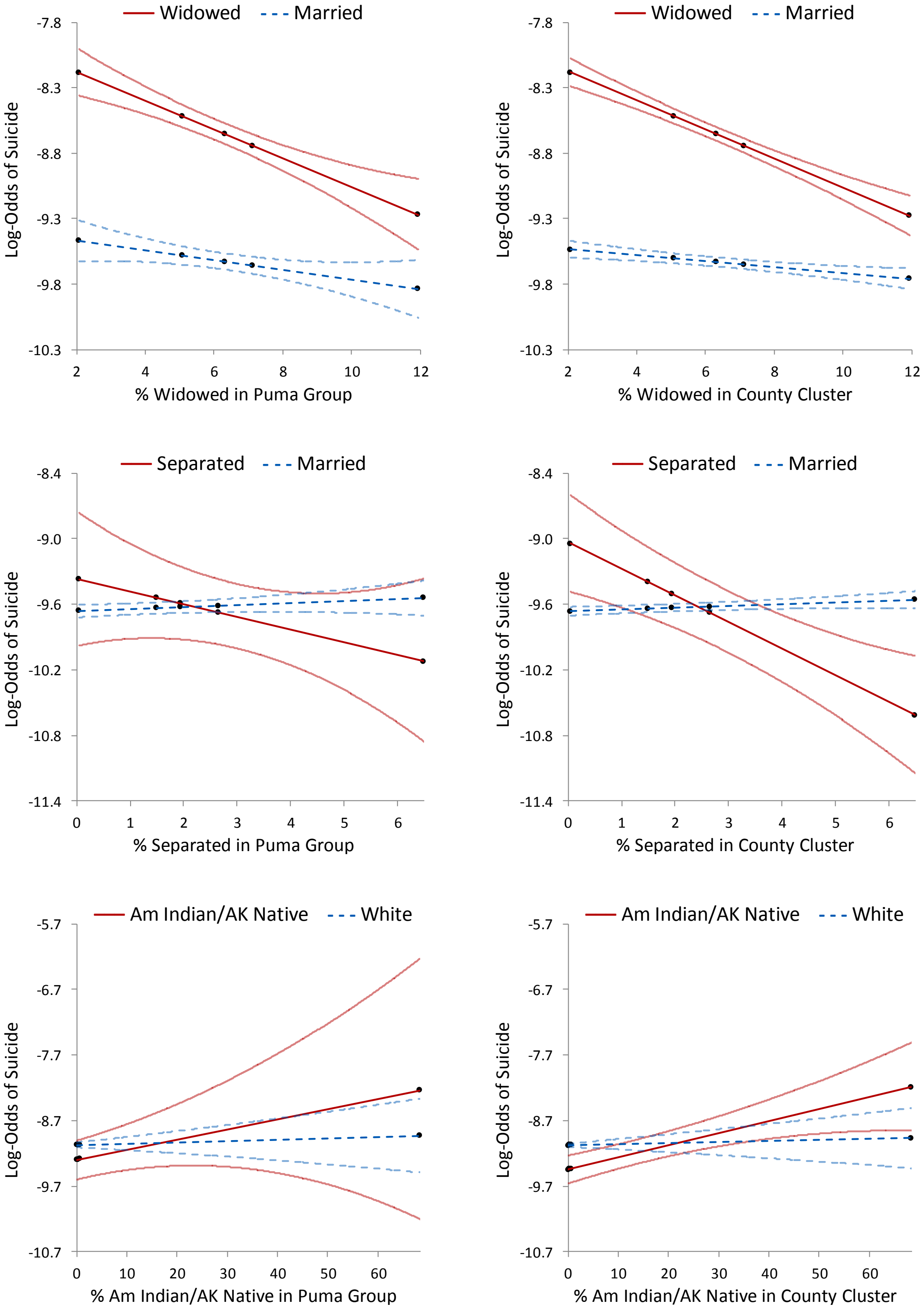


Figure 4. Comparison of logistic regression model effects plots for selected variable groups. Descriptive statistics used in all plots are based on Puma Groups. Confidence bands are consistent with a false discovery rate of .05. Dots on logit lines show the minimum, quartiles, median, and maximum values for the community rate variable (all but the maximum are less than 1 for American Indian/AK Native).

As noted above, CMF data provide county-level death and population counts by demographic group. These statistics enabled us to construct an alternative in-house database, similar in structure to the original but lacking the national origin and marital status variables. Within it, frequencies of cases and controls (population minus cases) for individual-level attribute profiles replaced weighted NVDRS and ACS records.

Performing logistic regressions on Puma Group– and County Cluster–based versions of this database,[9] we obtained effects plots for the AIAN predictor set resembling those in Figure 4. Exploring further, we adapted for use with each version a technique described by Hosmer, Lemeshow, and Sturdivant (2013) to evaluate data compatibility with logistic regression assumptions, replacing AIAN race and community rate predictors with dummy variables for AIAN and White race groups inside eight equal-length intervals of the rate variable range.

Logits for these groups, obtained from logistic regressions performed under this coding scheme, are presented in Figure 5. In each graph, a level number identifying the AIAN rate interval is plotted on the horizontal axis and the log-odds of suicide occurrence for observations within the group is plotted on the vertical axis. A table below each pair of graphs provides the midpoint for each level's interval, along with numbers of community–year combinations contributing data.

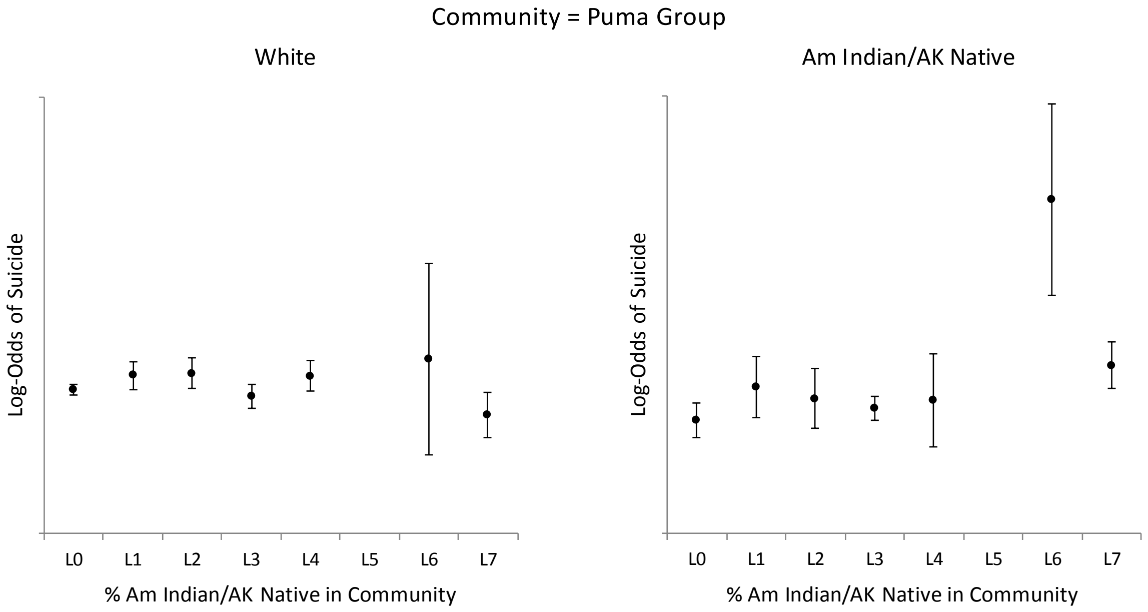
Differences between the Puma Group– and County Cluster–based graphs for AIAN race are striking. Logit height is nearly constant in the former, apart from substantial elevation at Level 6; but it rises almost monotonically (if not quite linearly) in the latter, showing substantial elevations beyond Level 4. Moreover, the rate range increases between the graphs, with the midpoint for Level 7 rising from 65% in the former to 91% in the latter.

Descriptive statistics presented in Table 3 furthermore show that AIAN population rates varied widely among County Cluster–year observation subsets within Puma Group–year rate levels. These findings collectively imply that the aggregation of counties to form Puma Groups did obscure distinctions crucial to detecting important effects.

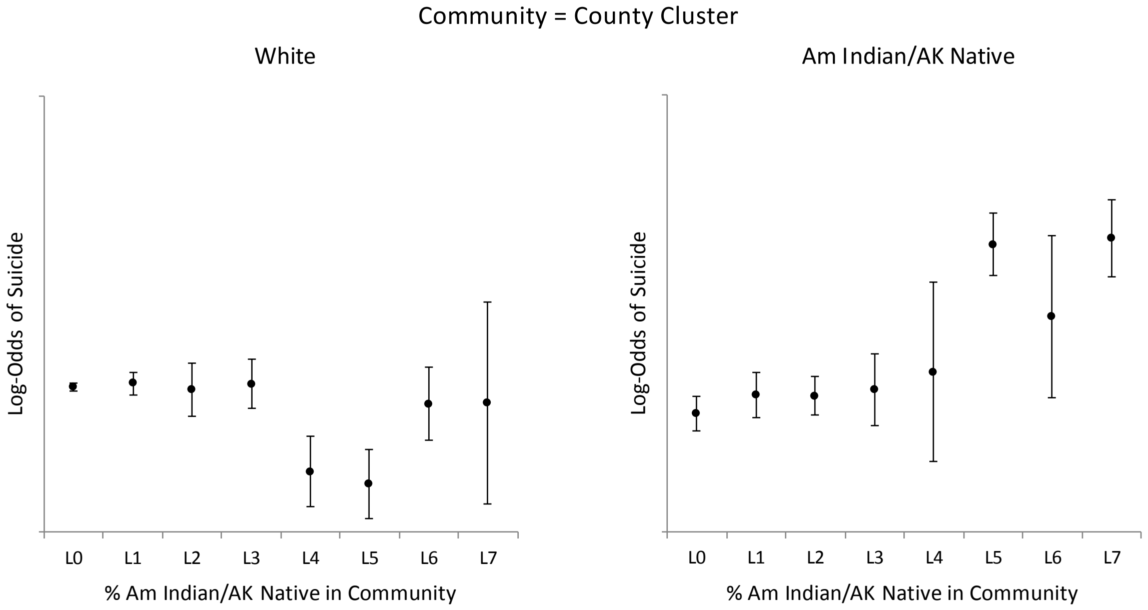
Table 3. Minimum and maximum American Indian/ Alaska Native population rates for County Cluster-Years within each rate level of Puma Group-Years (CMF-based reconstruction of 2005-2011 US-MSDS)

PG-Yr Rate Level	Level Mid- point	Num CC-Yrs	Min CC-Yr Rate	Max CC-Yr Rate
0	4	6,134	0	89
1	13	334	2	29
2	22	62	11	48
3	30	36	17	50
4	39	14	39	40
5	47	0	—	—
6	56	105	16	97
7	65	14	42	79

[9] For these and subsequently described models we used the Taylor series method of variance estimation.



Level	0	1	2	3	4	5	6	7	Total
Midpoint	4	13	22	30	39	47	56	65	
N PG-Yrs	1,913	66	10	6	14	0	7	7	2,023



Level	0	1	2	3	4	5	6	7	Total
Midpoint	6	18	30	42	54	66	79	91	
N CC-Yrs	6,225	301	46	44	17	5	35	26	6,699

Figure 5. Logit plots for suicide risk from CMF-based reconstruction of 2005-2011 US-MSDS. Confidence intervals are consistent with a false discovery rate of .05. Levels on the horizontal axis reflect equal subdivisions of the community rate variable's range. Markings on the vertical axis are suppressed for disclosure avoidance. *N PG-Yrs* and *N CC-Yrs* respectively indicate numbers of Puma Group–Year and County Cluster–Year combinations contributing observations to each level.

5. Discussion

This paper has described our effort to address longstanding problems in suicide risk factor research by constructing a novel database from multiple sources and geographic levels of information, including data furnished by the U.S. Census Bureau and the CDC. Our database design supported a suicide prediction model based on a multi-level, social network perspective that places suicide at the nexus of interacting individual and social factors.

We constructed in-house and laboratory versions of the database, which respectively combined public- and restricted-access ACS data with restricted-access NVDRS data. Theoretically relevant findings from statistical analyses conducted on the former version were amplified and extended when conducted on the latter, which suggests that increased geographic specificity in community definition afforded by the latter played an important role in detecting effects.

In what follows, we reflect on our experiences and their potential implications for further health-related research using multi-sourced, multi-level databases. After discussing database design, analytic strategy, and scientific findings, we consider the challenges and benefits of using public- and restricted-access ACS data.

5.1 Database Design, Statistical Analyses, and Scientific Findings

As noted above, our database was created by comingling individual-level case and control records, sourced respectively from NVDRS and ACS data sets, and attaching community-level variables sourced from ACS and other data sets. This construction allowed us to test hypotheses concerning the moderation of attribute-related individual risk by community attribute rates, thereby surmounting a bifurcation in analytic approaches that has constrained previous research.

Our approach was viable partly because the ACS person records themselves constituted a general population sample, while the NVDRS records comingled with them were few enough that they negligibly altered that sample's demographic composition. The rarity of suicide events that has typically hindered sample development thereby paradoxically facilitated our effort to develop a sample of cross-sectional design.

That design permitted statistical modeling of suicide as an outcome conditional upon multiple antecedents. Our logistic regression models predicted the log-odds of suicide occurrence for individuals having specific combinations of attributes within specific community circumstances, although their practical value is more likely to elucidate risk factor patterns.

The first of three such patterns presented in this paper concurs with theory-based predictions, regarding the notion that person-level factors normally thought to elevate suicide risk (in this case widowed marital status) are more readily tolerated where they are more prevalent. The second of these patterns (involving separated marital status) likewise concurs, to an extent that the attribute-associated risk apparently increases in some environments but decreases in others.

One might wonder how the third pattern, which involves American Indian/Alaska Native race, could be consistent with the same reasoning, given that suicide risk for AIAN status evidently *increases* with its community prevalence. A plausible explanation is that the models we tested relied upon community attribute rates to index supportiveness in the social environment. The elevations we observed in AIAN suicide risk were associated with high AIAN population rates. These rates may index residence in locations (e.g., Indian reservations) marked by degrees of physical, economic, and social hardship that outweigh the normally expectable support of shared circumstance (Mose, Bartholomew, and Weahkee, 2014).

A model refinement that might address this type of problem would be to characterize community environments with a richer set of community-level variables. We are, in fact, considering the inclusion of predictors that measure social capital (Lee and Kim, 2013) and physical and mental health levels (University of Wisconsin Population Health Institute, 2016) within communities.

A limitation less readily addressed is the relatively small number of individual-level NVDRS and ACS variables that were jointly available and compatibly codable.[10] Each of these data sets was designed to meet specific objectives, and the likelihood that either will be substantially expanded to include valuable information found only in the other seems low.

One response to this problem, however, might again draw benefit from the rarity of suicide. The demographic composition of the control group closely matches that of the community. One might therefore conduct analyses using only suicide cases for individual-level data records, predicting their attribute rates from community-level variables. Stated in terms of fourfold table analysis, this approach would estimate antecedent probability conditional upon outcome status, affording a perspective on risk that complements the one presented here while expanding the range of factors to examine.

5.2 Public- and Restricted-Access Data

Our approach to suicide risk factor research depends vitally upon data sets furnished by the U.S. Census Bureau and the CDC. These data sets exist through federally supported efforts to gather and disseminate sociodemographic, socioeconomic, and health-related information about the U.S. resident population. The agencies that do so must manage potential conflicts between the privacy needs of individuals and organizations on the one hand and the public interest benefits of scientific research on the other.

One means of addressing this challenge is to regulate the availability and use of sensitive data, which the U.S. Census Bureau and the CDC achieve partly through access restriction. To the best of our knowledge, this study marks the first project ever to combine restricted-access data from both these agencies within an RDC laboratory. Its implementation depended upon the willingness of management teams to cooperate across organizational boundaries in granting us the requisite permissions. We hope they view their efforts as having set a worthy precedent.

[10] We do expect in future publications to report findings for two additional variables we jointly coded, namely, unemployment status and physical problem presence.

A second means of addressing potential conflicts between privacy needs and public benefits in data use is the application of disclosure avoidance techniques when disseminating information publicly. The ACS PUMS files supporting our in-house database development, and in particular their use of PUMAs to identify residence location, exemplify this approach.

Our effort to surmount incompatibilities between PUMAs and counties through the introduction of PUMA Groups, when merging differently sourced datasets, met with mixed success. On the one hand, our laboratory-based analyses yielded effects missed by our in-house analyses for reasons plausibly attributable to the aggregation of diverse County Clusters within Puma Groups. On the other, our in-house analyses prepared us extensively for using laboratory resources.

Reviewing our efforts, we are considering alternative approaches to managing geographic unit conflict that could improve our use of ACS PUMS records. One of these could be heuristic assignment of county codes to these records, using the previously discussed Puma-county relationship files in conjunction with techniques such as multiple imputation.

In any case, both public- and restricted-access datasets clearly played a useful, and ultimately complementary, role in supporting our novel approach to suicide research. We hope that investigators studying similarly challenging subject matters (e.g., other types of low frequency illness, injury, or mortality) will agree with our conclusion that the integration of large scale sociodemographic, socioeconomic, and health-related datasets yields a whole that exceeds the sum of its parts.

6. References

Centers for Disease Control and Prevention. 2008. Strategic direction for the prevention of suicidal behavior: promoting individual, family, and community connectedness to prevent suicidal behavior. Atlanta, Georgia (http://www.cdc.gov/violenceprevention/pdf/Suicide_Strategic_Direction_Full_Version-a.pdf)

Curtin SC, Warner M, Hedegaard H. 2016. Increase in suicide in the United States, 1999–2014. NCHS data brief, no 241. Hyattsville, MD: National Center for Health Statistics

Durkheim E. 2006 [1897]. *On Suicide*. London: Penguin Classics

Fleiss JL. 1981. *Statistical Methods for Rates and Proportions, 2nd Edition.* New York: Wiley

Hanzlick R. 2007. *Death Investigations: Systems and Procedures.* Boca Raton, FL: CRC Press

Hosmer DW Jr, Lemeshow S, Sturdivant RX. 2013. *Applied Logistic Regression, 3rd Edition.* Hoboken, NJ: Wiley

Lauger A, Wisniewski B, McKenna L. 2014. Disclosure avoidance techniques at the U.S. Census Bureau: current practices and research. Center for Disclosure Avoidance Research, U.S. Census Bureau, Washington DC

Lee CJ, Kim D. 2013. A comparative analysis of the validity of US state- and county-level social capital measures and their associations with population health. *Social Indicators Research* 111:307–326

Mose AH, Bartholomew ML, Weahkee RL. 2014. Suicide mortality among American Indians and Alaska Natives, 1999–2009. *American Journal of Public Health* 104:S336–S342

National Center for Health Statistics (2015). Compressed Mortality File (Years 1999-2011, Series 2Q), as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program.

Pescosolido BA. 1994. Bringing Durkheim into the 21st century: A social network approach to unresolved issues in the study of suicide. In *Emile Durkheim: Le Suicide 100 Years Later*, ed. D Lester, pp. 264-295. Philadelphia: Charles Press

Pescosolido BA. 2011. Organizing the sociological landscape for the next decades of health and health care research: the Network Episode Model III-R as cartographic subfield guide. In *Handbook of the Sociology of Health, Illness, and Healing: A Blueprint for the 21st Century*, ed. BA Pescosolido, JK Martin, JD McLeod, A Rogers, pp. 39–66. New York: Springer

Pescosolido BA. 2012. Modeling multi-level influences on suicide. Application for federal assistance, Grant 11052784

Robinson WS. 1950. Ecological correlations and the behavior of individuals. *American Sociological Review* 15:351-357

Steensland B, Park JZ, Regnerus MD, Robinson LD, Wilcox WB, Woodberry RD. 2000. The measure of American religion: toward improving the state of the art. *Social Forces* 79:291-318

University of Wisconsin Population Health Institute. 2016. County health rankings and roadmaps. (http://www.countyhealthrankings.org/)

U.S. Census Bureau. 2007. *County and City Data Book: 2007 (14th edition).* Washington, DC

U.S. Census Bureau. 2009. *A compass for understanding and using American Community Survey data: what PUMS data users need to know.* Washington, DC

U.S. Census Bureau. 2010. Substantial changes to counties and county equivalent entities: 1970-present. Washington, DC (https://www.census.gov/geo/reference/county-changes.html)

U.S. Department of Health and Human Services (HHS) Office of the Surgeon General and National Action Alliance for Suicide Prevention. 2012. National strategy for suicide prevention: goals and objectives for action. Washington, DC

Wray M, Colen C, Pescosolido BA. 2011. The sociology of suicide. *Annual Review of Sociology* 37:505-528