Paper 1951-2018

Development of an Individual Level Social Determinants of Health (SDoH) Model

Ryan C. Butterfield, DrPH, Melissa Gottschalk, MPH, Paul A. LaBrec, MS, 3M Health Information Systems

ABSTRACT

Growing evidence in the health services literature suggests that social determinants of health (SDoH) can affect a person's risk for adverse health events and increased cost of care. Individuals can be scored on those factors associated to SDoH measures that are individual in nature and weights can be created that can be used for generating greater insight into risk, costs, and health care utilization patterns.

To develop this prototype model, data elements will consist of Z Codes from the claims data of a large Midwest Medicaid payer. Z Codes were created as a new set of supplemental codes with the advent of ICD-10 and are physician-coded attributes at a person level. We examine the subset of Z Codes (55-65) which indicate "Persons with potential health hazards related to socioeconomic and psychosocial circumstances." These are set as binary indicators at the individual subject level. A subset of the Z Codes is selected for use in this analysis.

SAS 9.4 will be used to research and develop this procedure. Initial research indicates the use of a Latent Class statistical model as best option for creating the classification model. This model will be developed and tested for optimization, feasibility, and accuracy. From this model predicted probabilities will be output which can then be used in a generalized linear regression model as weights to help define, adjust, and assess the outcomes of interest and combine with 3M Clinical Risk Group (CRG) scores as a model covariate.

INTRODUCTION

The Centers for Disease Control and Prevention defines social determinants of health (SDoH) broadly as "Conditions in the places where people live, learn, work, and play." Adverse social determinants include unstable housing, low income, unsafe neighborhoods, or substandard education (CDC, 2018). There is a large and growing literature demonstrating the impact of social determinants on health status and various health outcomes (Healthy People 2020).

An important challenge in accounting for social determinants of health in health services research is the measurement of social determinants. Many determinants—income, education, housing, literacy, insurance status, food security and others—are measured in the national census and other survey datasets and are available for analysis in public datasets summarized by small areas (e.g., Census Tracts). The authors summarized the development of such a geographically-based score in a previous conference paper (LaBrec and Butterfield, 2017).

With advent of the International Classification of Diseases Tenth Revision (ICD-10) a new category of supplemental codes labeled 'Z Codes' was created. In the ICD-10 classification scheme, Z Codes are found in Chapter 21, "Factors influencing health status and contact with health services (Z00-Z99)" (CMS, 2016). Among these new "Z" codes is the following series related to potential hazards due to family and social circumstances impacting health status:

Z55-Z65 - Persons with potential health hazards related to socioeconomic and psychosocial circumstances (ICD-10 Data.com, 2016).

- Z55 - Problems related to education and literacy
- Z56 - Problems related to employment and unemployment
- Z57 - Occupational exposure to risk factors
- Z59 - Problems related to housing and economic circumstances

Z60 - Problems related to social environment

Z62 - Problems related to upbringing

Z63 - Other problems related to primary support group, including family circumstances

Z64 - Problems related to certain psychosocial circumstances

Z65 - Problems related to other psychosocial circumstances

Each of these codes has sub-codes providing a more specific description of the problem. Some of these codes describe issues traditionally recognized as related to socioeconomic status:

Z59 - Problems related to housing and economic circumstances

Z59.0 - Homelessness

Z59.1 - Inadequate housing

Z59.4 - Lack of adequate food and safe drinking water

Z59.5 - Extreme poverty

Z59.6 - Low income

Z59.7 - Insufficient social insurance and welfare support

While others are not traditional measures of social factors:

Z60.2 - Problems related to living alone

Z60.3 - Acculturation difficulty

Z60.5 - Target of (perceived) adverse discrimination and persecution

Z63.1 - Problems in relationship with in-laws

Z62.1 - Parental overprotection

An advantage of using Z Codes over a geographically-defined measurement of SDoH using aggregate data is that the codes are specific to individuals and included on health care claims, along with other health information. As these codes are specific to individuals they can potentially be used in supplementing clinical risk adjustment with risk adjustment for social factors. A disadvantage of using Z Codes is that the prevalence of Z Codes relating to social determinants of health on medical records is currently low. This paper presents one approach to defining categories of persons based on Z Codes appearing on their health care claims.

METHODOLOGY

The claims database used in this analysis is from a Midwest Medicaid Insurance Payer and spans from October 2015 to March 2016. The Z Codes are collected from the individual or medical provider as part of the medical interview and are part of the medical record and claims data file. Those Z Codes from categories 55-65 are part of the individual assessments of social risk factors. Domain level definitions of the ICD-10 Z Codes are used in this analysis can be found in Table 1.

Z55 Problems related to education and literacy
Z56 Problems related to employment and unemployment
Z57 Occupational exposure to risk factors
Z59 Problems related to housing and economic circumstances
Z60 Problems related to social environment
Z62 Problems related to upbringing

Z63 Other problems related to primary support group, including family circumstances
Z64 Problems related to certain psychosocial circumstances
Z65 Problems related to other psychosocial circumstances

Table 1. Domain Level Definitions of Z Codes Used in the Analysis

STATISTICAL ANALYSIS

Data are described using traditional descriptive statistics. Domain identification and development is done using Latent Class Analysis. This analysis used PROC LCA in SAS v9.4, the Z Codes from the claims data are analyzed for item-response patterns that may be defined to be related to the SDoH domains. Segmentation profiles representing Social Determinants of Health domains as based on the Z Codes are produced for individual level data which can allow for identifying those individuals having a greater probability of being part of a given domain/class. Prevalence of claims for the classes are also computed as part of this analysis.

Lanza and Collins popularized the latent modeling through their applications to social and clinical behaviors (Lanza and Collins, 2010; Lanza et al, 2011). They also changed the availability of performing Latent Class Analysis (LCA) by developing the LCA procedure for SAS (Lanza et al, 2007; Lanza et al, 2011). These procedures had been available via other software platforms, but the LCA procedure did allow the SAS platform to be used for model development and estimation. The use of LCA has gained traction, with more research using the technique for clustering of categorical data. This technique sits with its counterparts: Factor Analysis, Latent Trait Analysis, and Latent Profile Analysis in the manner presented in Table 2.

Comparison of Clustering Methods		**Manifest (Observed) Variables**	
		Metrical	Categorical
Latent Variables	Metrical	**Factor Analysis**	**Latent Trait Analysis**
	Categorical	**Latent Profile Analysis**	**Latent Class Analysis**

Table 2. Comparison of Clustering Methods

Market segmentation or patient profiles are just two examples of applications of LCA that have become widely used. We will use Lanza, Collins, Lemmon, et al. (2007) and Collins and Lanza (2011) for consistency of nomenclature to outline the basic latent class model as the following. Let K be latent subgroups drawn from nominal variables with j = 1,…,J observed variables, and j has rj = 1, …, Rj response categories. Let x = (r1,…rh) be the vector of item responses per individual, J, being observed from the claims data. Let the latent variable be represented by C which has the latent classes c = 1,…,K. Given this let an indicator function be identified that will allow for dichotomous response for the latent variable C. This indicator function can be defined as I(xj,…,rj) which allows for the function to equate to 1 when j = rj and otherwise equals 0. Thus, defining the 1=yes and 0=no definitions of a classical approach to dichotomous variables. Therefore, the probability of observing any given response pattern is given by the statistical Latent Class Model:

$$\Pr\{X = x\} = \sum_{c=1}^{K} \gamma_c \prod_{j=1}^{J} \prod_{r_j=1}^{R_J} \rho_{j,r_j|c}^{I(x_j=r_j)}$$

Where membership probability or prevalence in the latent class is denoted by $\gamma_c$and the item response probability for rj is given by $\rho_{j,r_j|c}^{I(x_j=r_j)}$to item j in latent class c. This model can be extended to include covariates, but that is beyond the scope of this paper.

When considering the number of classes in an LCA model the model fit methodology chosen is vitally important, as is the case with any statistical model. With LCA, however, it can be somewhat vague as which model fit methodology is optimal. The available options from the LCA procedure in SAS are entropy, G-squared, AIC, BIC, CAIC, adjusted BIC, and log-likelihood (Lanza et al, 2007; Lanza et al, 2011) Larsen et al. (2017) discuss the various pros and cons concerning the model fit statistics specifically for LCA. Finch (2015) compares the model fit statistics through a Monte Carlo simulation that is just for information criteria and but does not include the entropy measure. These findings indicate that the adjusted Bayesian Information Criteria (aBIC/Adjusted BIC) performed better than the others. The Bootstrapped Likelihood Ratio Test, however, does have indication for being the most robust model fit tool available (Finch, 2015; Dziak et al., 2011). The LCA model is optimized when it operates under the assumption of local independence. This assumption is that manifest or measures are independent of each other within latent classes (Lanza et al, 2012).

To compare the latent classes and identify a possible indication of their effect, we used two primary outcome measures from claims data for this analysis. These include Total Medical Allowed dollars (TMA), and Outpatient ER Allowed dollars. Total Medical Allowed = inpatient + outpatient + professional claim dollars. Outpatient ER is a subset of Outpatient dollars. Clinical risk was represented by a weight and calculated using a combination of 3M's Clinical Risk Group™ (CRG), age group, and sex of each patient (Averill et al., 1999). Weights are monotonic in that increased weight = increased clinical severity.

DATA SUMMARY

The prevalence of claims with at least one Z Code in the range 55.x through 65.x in this dataset account for roughly 0.67% of the total claims data. The resulting total available sample of Z Code data is 15,753 claims from the complete dataset of 2,341,310 claims containing diagnoses (Table 3). Slightly more than half the patients (55.4%) were male. Males in the dataset were younger than females (17.4 vs 20.3 years) (Table 4).

Z Code	Domain	N	%
Z55	Problems related to education and literacy	1786	11.3
Z56	Problems related to employment and unemployment	202	1.3
Z57	Occupational exposure to risk factors	35	0.2
Z59	Problems related to housing and economic circumstances	1769	11.2
Z60	Problems related to social environment	859	5.5
Z62	Problems related to upbringing	5953	37.8
Z63	Other problems related to primary support group, including family circumstances	3713	23.6
Z64	Problems related to certain psychosocial circumstances	185	1.2
Z65	Problems related to other psychosocial circumstances	1250	7.9
	TOTAL	1572	100.0

Table 3. Frequency of Z Code Use in Claims

Gender	N	%	Mean Age	Std Dev	Minimum	Maximum
Female	7029	44.6	20.3	16.8	0	99.0
Male	8724	55.4	17.4	15.1	0	95.0

Table 4. Patient Demographics

SELECTING A MODEL IN LATENT CLASS ANALYSIS

A focus of this paper is discussing model fit statistics including entropy, G-squared, AIC, BIC, CAIC, adjusted BIC, and log-likelihood. These are presented and discussed as tools for identifying the optimal number of classes in the LCA model. Once a model is decided upon, subjects were assigned to classes based on the highest item response probability per item, which defines their assignment to a single class. Once subjects are assigned to classes then they can be associated to covariates of interest such as demographics, health outcomes, or health costs. Using the LCA procedure, multiple latent class models are evaluated. This is a *brute force approach* to model fitting. Alternatives include using a Bootstrapped Likelihood Ratio test (Dziak and Lanza, 2016). This approach, however, requires the IML procedure in SAS which may not always be readily available to users.

The basic LCA procedure follows, using the Z Code data for this example:

```
  options linesize=200;
PROC LCA data=newsas.test_miss_2 OUTSEEDS=ID OUTPARAM=param OUTSTDERR=stderr1;
nclass 6;
items Zz550 Zz553 Zz554 Zz558 Zz559
…
Zz658 Zz659;
categories
2 2 2 2 2
…
2 2
;
seed 1313;
nstarts 20;
maxiter 10000;
RHO PRIOR = 1;
Id person_id;
run;
```

A few notes on the procedure code, as the number of *items* must have a 1:1 correspondence with the *categories*. The RHO prior is a required statement to be able to calculate the standard error of the classes. The *OUT*- commands allow for various model information to be captured including the item response probabilities and different model estimates. The *seed* should be set as constant when doing this brute force approach so the model has a constant starting position.

Recommended best practice in LCA is to run multiple models while varying the class size for the purpose of identifying the most robust model fit. These model fit statistics include traditional model fit measures based on information criteria, but also measures like entropy. For LCA Collins and Lanza suggest minimizing the entropy measure will lead to the best fit and number of classes, while the other measures are interpreted in their traditional manner. Models with numbers of classes from 3 to 10 were run using the LCA procedure, with the 6-class model being used for discussion (Table 5).

Latent Classes Modeled	Log Likelihood	G Squared	AIC	BIC	CAIC	Adjusted BIC	Entropy
3	13792.67	4441.22	4817.22	5937.56	6125.56	5340.22	0.99
4	13623.90	4103.68	4605.68	6101.46	6352.46	5303.94	0.99
5	13547.14	3950.16	4578.16	6449.38	6763.38	5451.69	0.93
6	**13435.83**	**3727.54**	**4481.54**	**6728.19**	**7105.19**	**5530.32**	**0.85**
7	13369.96	3595.80	4475.80	7097.88	7537.88	5699.84	0.93
8	13226.55	3308.98	4314.98	7312.50	7815.50	5714.29	0.96
9	13082.16	3020.20	4152.20	7525.15	8091.15	5726.76	0.99
10	13019.91	2895.70	4153.70	7902.08	8531.08	5903.53	0.96

Table 5. Fit Statistics Summary by Class Model

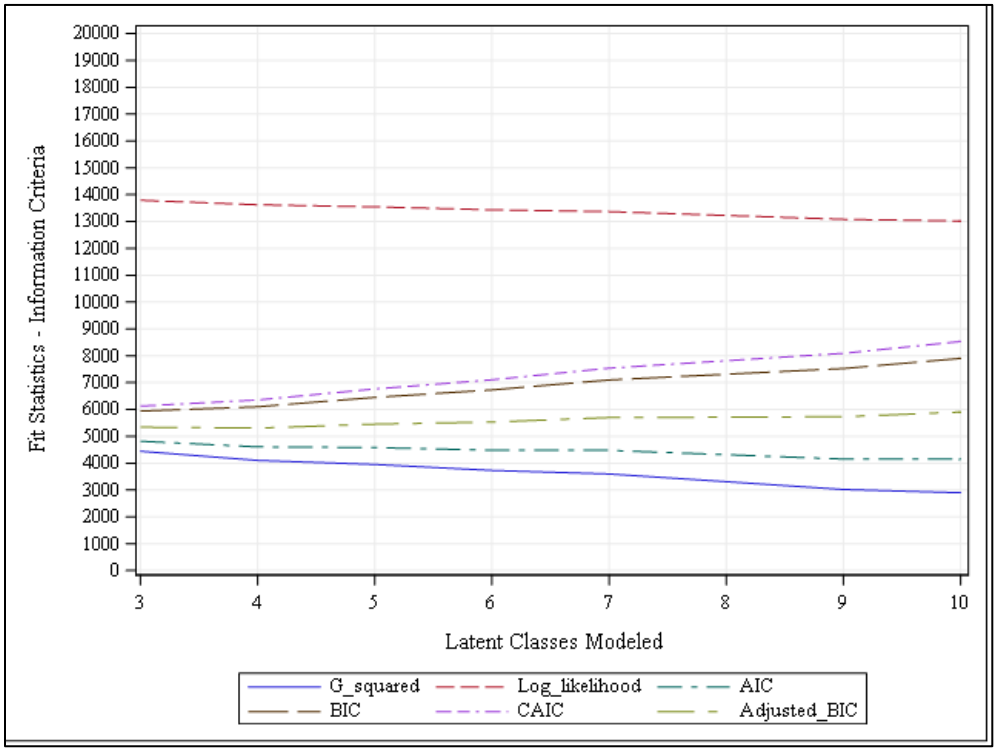

Figure 1. Fit Statistics - Information Criteria

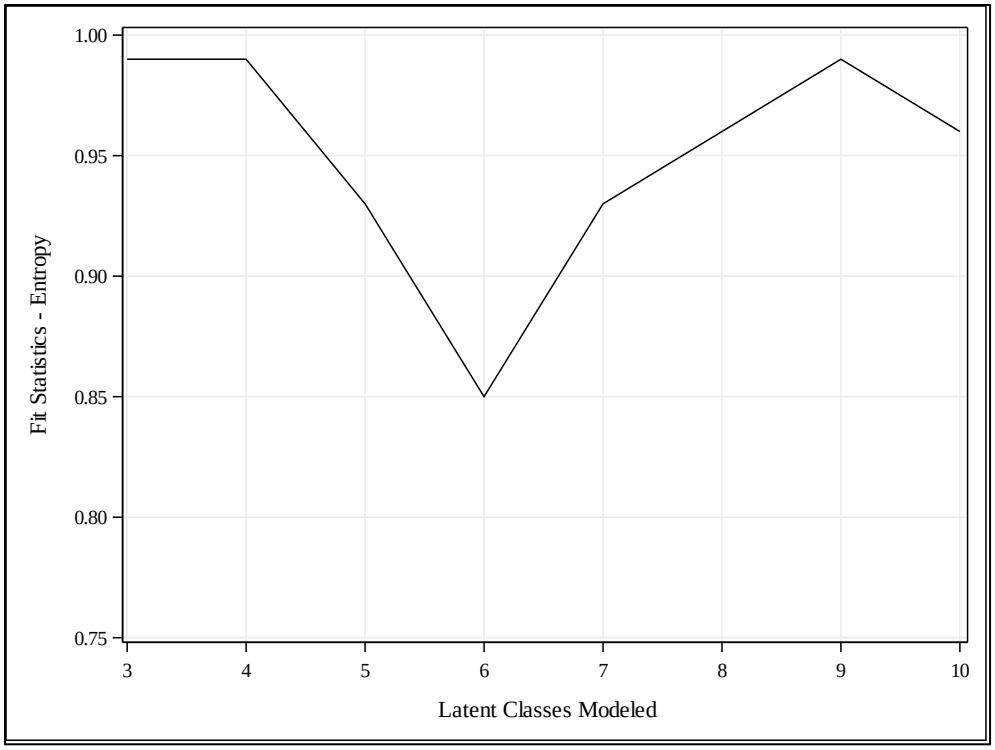

Figure 2. Fit Statistics - Entropy

Overall it appears that the model selection, depending on the fit statistic, favors either a lower class model (AIC, BIC) or a much higher class model (log-likelihood), however the model of size 6 is selected as it would appear to be a middle ground using the aforementioned fit measures, as well as the entropy score to inform the decision. The size 6 model indicates the lowest entropy which would be the preferable state of information being used.

CLASS MEMBERSHIP AND COMPARING CLAIMS DATA FOR Z CODE-BASED LATENT CLASSES

The latent class is the underlying "unobserved" phenomena that helps to explain why the manifest variables are clustered together. The IR is the item response for that class given the variable and is the conditional probability of response for that intersection. The claims outcomes are linked for comparison of the classes. Total Medical and Outpatient ER Allowed are measured in dollars while the disease severity weight is a continuous measure.

A variable is defined to be conditionally in a class when the item response probability (Rho) is highest. It is often useful to use indicator variables with the *max* command to assign a variable to its dominant class. For example, if it is identified that the optimal number of classes is 6, then there may be up to 6 item response probabilities for a single variable. For example, consider the following results from the 6-class model (Table 6). The situation for latent class 3 is interesting to examine more closely. When assigning the variables to clusters, clearly class 3 is not being indicated. This is due to there not being any maximum probabilities found. This indicates that while class 3 may have similarities to others it serves as a remainder class where those subjects who didn't clearly align to the other classes.

Latent Class Cluster	Z Code Description	Latent Class 1 IR	Latent Class 2 IR	Latent Class 3 IR	Latent Class 4 IR	Latent Class 5 IR	Latent Class 6 IR	Total Medical Allowed ($)	Outpatient ER Allowed ($)	Disease Severity Weight
1	Acculturation difficulty	0.057760	0.025860	0.000012	0.000007	0.000007	0.000010	7670.44	300.45	1.877

Latent Class Cluster	Z Code Description	Latent Class 1 IR	Latent Class 2 IR	Latent Class 3 IR	Latent Class 4 IR	Latent Class 5 IR	Latent Class 6 IR	Total Medical Allowed ($)	Outpatient ER Allowed ($)	Disease Severity Weight
1 (cont)	Exposure to disaster, war and other hostilities	0.002506	0.000000	0.000000	0.000000	0.000000	0.000000	5584.26	953.27	4.016
	Inadequate housing	0.009862	0.003916	0.000006	0.008683	0.008514	0.000004	9236.40	385.76	3.664
	Problem related to upbringing, unspecified	0.002453	0.000002	0.000002	0.002381	0.000000	0.000001	13533.74	53.33	1.123
	Problems related to education and literacy, unspecified	0.137913	0.122018	0.000066	0.043652	0.002863	0.005132	9029.94	242.067	1.962
	Target of (perceived) adverse discrimination and persecution	0.002506	0.000000	0.000000	0.000000	0.000000	0.000000	6185.02	34.95	2.115
2	Educational maladjustment and discord with teachers and classmates	0.015071	0.027337	0.000011	0.000014	0.000004	0.000006	8142.32	216.04	1.961
	Extreme poverty	0.000001	0.005762	0.000001	0.000001	0.000001	0.000001	7251.17	164.60	4.350
	Illiteracy and low-level literacy	0.000001	0.005757	0.000002	0.000004	0.000001	0.000001	6142.30	299.62	2.399
	Other physical and mental strain related to work	0.000002	0.015153	0.000007	0.001816	0.000002	0.007823	6688.11	1429.76	2.194
	Other problems related to education and literacy	0.004972	0.065481	0.006399	0.006588	0.000010	0.011768	4715.81	156.24	1.402
	Other problems related to social environment	0.000004	0.022058	0.005399	0.005969	0.002818	0.003626	28873.84	355.67	5.308
	Other specified problems related to psychosocial circumstances	0.000023	0.207502	0.000057	0.010718	0.005653	0.007671	28687.50	495.14	6.998
	Problem related to housing and economic circumstances, unspecified	0.000007	0.055081	0.000014	0.000014	0.017122	0.007084	6489.90	531.83	3.395
	Problem related to primary support group, unspecified	0.032558	0.254512	0.000255	0.000047	0.011444	0.017161	7296.56	426.67	2.518
	Problem related to social environment, unspecified	0.007527	0.085905	0.016587	0.000022	0.000011	0.000022	12454.69	533.08	3.716

Latent Class Cluster	Z Code Description	Latent Class 1 IR	Latent Class 2 IR	Latent Class 3 IR	Latent Class 4 IR	Latent Class 5 IR	Latent Class 6 IR	Total Medical Allowed ($)	Outpatient ER Allowed ($)	Disease Severity Weight
2 (cont)	Problem related to unspecified psychosocial circumstances	0.000007	0.061225	0.013910	0.000015	0.000008	0.000011	8902.56	622.22	3.422
	Social exclusion and rejection	0.000001	0.010080	0.000003	0.000003	0.000001	0.000002	32734.01	299.23	9.106
	Underachievement in school	0.017519	0.098252	0.000024	0.000019	0.000013	0.007254	3186.29	308.66	1.471
4	Dependent relative needing care at home	0.000000	0.000001	0.000001	0.002357	0.000000	0.000000	5712.66	858.01	4.127
	Disappearance and death of family member	0.001708	0.000044	0.000039	0.111163	0.000016	0.016162	9921.12	637.73	2.462
	Discord with neighbors, lodgers and landlord	0.000000	0.000001	0.000001	0.001178	0.000000	0.000000	6174.30	565.68	4.648
	Disruption of family by separation and divorce	0.009449	0.015559	0.000051	0.088300	0.000016	0.010231	6463.34	319.84	1.638
	Imprisonment and other incarceration	0.000002	0.000006	0.007709	0.013568	0.000002	0.000003	8766.07	882.07	3.486
	Inadequate parental supervision and control	0.000000	0.000001	0.000001	0.002357	0.000000	0.000000	2586.36	0.00	1.347
	Insufficient social insurance and welfare support	0.000000	0.000001	0.000001	0.001178	0.000000	0.000000	12280.74	457.36	2.259
	Lack of adequate food and safe drinking water	0.000000	0.000001	0.000001	0.002357	0.000000	0.000000	8205.32	1665.66	3.113
	Low income	0.017514	0.000006	0.004786	0.022952	0.000005	0.008066	15396.81	440.37	2.474
	Occupational exposure to dust	0.000000	0.000001	0.000001	0.002357	0.000000	0.000000	7618.86	357.24	5.969
	Occupational exposure to other risk factors	0.000001	0.000003	0.000002	0.004713	0.000001	0.000001	4233.03	782.52	0.913
	Occupational exposure to unspecified risk factor	0.000000	0.000001	0.000001	0.001178	0.000000	0.000000	8494.32	942.62	2.647
	Other problems related to housing and economic circumstances	0.000004	0.012234	0.000014	0.023123	0.002700	0.003831	9259.68	352.34	4.230
	Parental overprotection	0.000001	0.000001	0.000002	0.005894	0.000001	0.000001	2409.66	41.33	0.939
	Problems in relationship with spouse or partner	0.004431	0.017465	0.000034	0.084095	0.002250	0.000020	6224.48	909.42	2.188

Latent Class Cluster	Z Code Description	Latent Class 1 IR	Latent Class 2 IR	Latent Class 3 IR	Latent Class 4 IR	Latent Class 5 IR	Latent Class 6 IR	Total Medical Allowed ($)	Outpatient ER Allowed ($)	Disease Severity Weight
4 (cont)	Problems of adjustment to life-cycle transitions	0.004995	0.000003	0.000002	0.007077	0.000001	0.004044	12411.45	691.29	3.103
	Problems related to living alone	0.000005	0.007497	0.000015	0.031590	0.000005	0.000008	24999.06	870.18	9.248
	Problems related to living in residential institution	0.000004	0.000014	0.000012	0.027180	0.002672	0.000006	53335.02	134.21	11.598
	Problems related to multiparity	0.000011	0.000031	0.000033	0.088395	0.000012	0.000018	6580.69	543.88	2.114
	Problems related to other legal circumstances	0.008071	0.005940	0.000017	0.038625	0.008292	0.016107	13904.52	261.90	2.002
	Problems related to unwanted pregnancy	0.000002	0.000007	0.000007	0.017678	0.000002	0.000004	4972.08	1344.42	1.735
	Unemployment, unspecified	0.000007	0.014843	0.000015	0.031852	0.027629	0.000011	12917.36	1035.73	4.333
	Unspecified problems related to employment	0.000001	0.003567	0.000003	0.006513	0.000001	0.000002	6481.78	181.78	1.873
	Victim of crime and terrorism	0.000002	0.000005	0.000006	0.016501	0.000002	0.000003	9765.45	1716.46	2.714
5	Homelessness	0.000052	0.000080	0.000098	0.007889	0.999432	0.000134	19674.72	1961.76	5.094
6	Other specified problems related to primary support group	0.064988	0.014805	0.012545	0.000071	0.005721	0.998603	7801.65	447.99	2.359

Table 6. 6-Class LCA Model Results

Clearly there is variable response that aligns with multiple classes, but given we are dealing with probability, then the maximum is a sufficient decision tool for assigning certain variables to classes. We can see how costs and severities change with response to the presence of different Z Codes. Looking at these same outcomes with regards to the latent classes allows for a higher level of insight to be evaluated. Looking at the summary of the outcomes by Latent Class, we can see how there is potentially an underlying influence on these outcomes given the grouping of the Z Codes into these domains (Table 7).

There is indication that the different latent classes do have variance from each other with regards to our outcomes. Class 5 (homelessness) showing the highest disease severity, ER Allowed dollars, and Total Medical Allowed dollars. Using these clusters to examine the outcomes allows for an easier evaluation of the effect of having these manifest variables present and to the extent that they present as determinants of health, either positive or negative.

Latent Class Grouping	Total Medical Allowed ($)	Outpatient ER Allowed ($)	Disease Severity Weight
1	8539.97	328.30	2.460
2	12428.09	449.13	3.711
4	10796.42	666.34	3.382
5	19674.72	1961.76	5.094
6	7801.65	447.99	2.359

Table 7. Allowed Dollars by Latent Class Results

DISCUSSION

The impact of social determinants of health on health status, health services utilization, and health care costs has generated increased study, policy concern, and the attention of both health care providers and payers over the past few years. For example, some authors have shown that state Medicaid agencies can spend as much as $3 on social services for every $1 spent on health benefits for their beneficiaries (Bradley et al, 2016). A component of the value-based payment debate has focused on the question of adjusting health care outcomes evaluation and payment not only for clinical risk, but for social risk factors as well (National Quality Forum, 2014).

Challenges in measuring SDoH at either a neighborhood or individual level are numerous and Healthy People 2020 is a testament to that difficulty. The use of public data by small geographies has many advantages, including availability, standardization, and low cost, but is still ultimately a proxy for information on a specific individual. The advent of the ICD-10 classification system has presented a claims-based source of information on the social and psychosocial determinants of health previously not available.

As our understanding of the effect of SDoH on health outcomes becomes clearer, there is a need for more statistically sophisticated methods to garner insights into the complex interactions that impact human health. There are several challenges in *modeling* social determinants of health. The main being that each individual member of the population has unique determinants of health and this variance serves to dilute patterns that could be easily detected through traditional statistical methodologies. More cluster related methods like Latent Class Analysis (LCA) and Principal Component Analysis (PCA) are valuable in identifying patterns as they sort through the noise to identify the underlying latent variables that may be underlyingly affecting the patterns in the data. Additionally, when pursuing the goal of an individual level score, using Z Codes as a sole data source may be hindering for two reasons. First, the low prevalence of the presence of Z Codes in claims data is hindering to any statistically generalizable approaches. Second, not all Z Codes in the range of social and psychosocial codes are specific to social determinants, they also include psycho-social and environmental determinants. Future research, including input from social scientists, is needed to refine the categorical definition of social determinants from health care claims data.

This work has promise in the refinement of clinical risk adjustment for social factors impacting health and the utilization of and payment for health services. Health care providers have long been aware that they treat patients who may be clinically very similar, but who live in quite different social environments. One 60-year-old male with multiple chronic conditions may be released from the hospital after surgery into an environment with stable housing, food security, adequate financial resources, ability to get to medical appointments, and a family who can advocate for the patient's well-being. Another patient in the same demographic and clinical risk group may be released from the same hospital after the same surgery into an environment of homelessness, instability, and lack of a social support network. The outcomes and costs for the two individuals can be quite different and clinical risk-adjustment alone won't discern these differences.

In addition to claims-based sources of data on social determinants of health, we are investigating other sources including patient-reported outcomes (PRO) surveys. Some of these include questions regarding the social situation of patients and can be used to measure social determinants.

CONCLUSION

The impact of social determinants of health on health status, health services utilization, and health care costs has gained increased study, policy concern, and the attention of both health care providers and payers over the past few years. In this paper, we have presented a potential method for creating categorical measurement of SDoH variables for use in health services research and payment risk-adjustment. This approach is meant to contribute to the identification of available sources of data on social determinants of health and the advancement of analytic methods to assess their impact on health and health care utilization and payment.

REFERENCES

Akaike H. A new look at the statistical model identification. IEEE Trans Autom Control. 1974 Dec;19(6):716–23.

Averill, R et al. Development and Evaluation of Clinical Risk Groups (CRGs). 3M HIS Research Report 9-99, 1999. Accessed March 6, 2017. http://solutions.3m.com/3MContentRetrievalAPI/BlobServlet?lmd=1225920653000&assetId=1180606514454&assetType=MMM_Image&blobAttribute=ImageFile

Bartholomew DJ, Knott M, Moustaki I. Latent variable models and factor analysis: a unified approach. 3rd edition. West Sussex, UK: John Wiley & Sons; 2011.

Bentler PM, Bonett DG. Significance tests and goodness of fit in the analysis of covariance structures. Psychol Bull. 1980;88(3):588–606.

Bradley EH et al. Variation In Health Outcomes: The Role Of Spending On Social Services, Public Health, And Health Care, 2000-09. Health Affairs 35, no.5 (2016):760-768.

Centers for Disease Control and Prevention. Social Determinants of Health: Know What Affects Health. Accessed on 7-March-18 at https://www.cdc.gov/socialdeterminants/

Centers for Medicare and Medicaid Services. ICD-10-CM Official Guidelines for Coding and Reporting: FY 2016. Accessed November 15, 2016 at http://www.cdc.gov/nchs/data/icd/10cmguidelines_2016_final.pdf

Collins LM, Lanza ST. Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences. New York, NY: Wiley; 2010.

Dayton CM. Latent class scaling analysis. Quantitative Applications in the Social Sciences. Thousand Oaks: Sage; 1999.

Dziak, J. J., & Lanza, S. T. (2016). LCABootstrap SAS macro users' guide (version 4.0). University Park: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu

Finch, H. (2015) A Comparison of Statistics for Assessing Model Invariance in Latent Class Analysis. Open Journal of Statistics, 5, 191-210. http://dx.doi.org/10.4236/ojs.2015.53022

Green MJ. Latent class analysis was accurate but sensitive in data simulations. J Clin Epidemiol. 2014 Oct;67(10):1157–62. pmid:24954741

Healthy People 2020 [Internet]. Washington, DC: U.S. Department of Health and Human Services, Office of Disease Prevention and Health Promotion [cited February 28, 2017]. Available from: http://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-health

ICD-10 Data.com. Factors influencing health status and contact with health services Z00-Z99. Accessed November 15, 2016 at http://www.icd10data.com/ICD10CM/Codes/Z00-Z99

LaBrec P and Butterfield R. The Development and Application of a Composite Score for Social Determinants of Health. SAS Global Forum 2017 Proceedings. Orlando FL, April 2017.

Lanza ST, Collins LM, Lemmon D, Schafer JL. PROC LCA: a SAS procedure for latent class analysis. Struct Equ Modeling. 2007;14(4):671–694. [PMC free article] [PubMed]

Lanza ST, Dziak JJ, Huang L, Xu S, Collins LM. PROC LCA & LTA User's Guide Version 1.2.7. University Park: Pennsylvania State University; 2011.

Lanza ST, Flaherty BP, Collins LM. Latent class and latent transition models. In: Schinka JA, Velicer WF, editors. Research Methods in Psychology. Hoboken, NJ: Wiley; 2012.Handbook of Health Psychology; vol 2, 2nd ed.

Larsen FB, Pedersen MH, Friis K, Glümer C, Lasgaard M (2017) A Latent Class Analysis of Multimorbidity and the Relationship to Socio-Demographic Factors and Health-Related Quality of Life. A National Population-Based Study of 162,283 Danish Adults. PLoS ONE 12(1): e0169426. doi:10.1371/journal.pone.0169426

National Quality Forum. Risk Adjustment for Socioeconomic Status or Other Sociodemographic Factors: Technical Report. August 15, 2014. Accessed June 9, 2016. http://www.qualityforum.org/Risk_Adjustment_SES.aspx

Nylund KL, Asparouhov T, Muthen BO. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Struct Equ Model Multidiscip J. 2007 Oct;14(4):535–69.

ACKNOWLEDGMENTS

We would like to thank our team and colleagues at 3M Health Information Systems.

RECOMMENDED READING

- Collins LM, Lanza ST. Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences. New York, NY: Wiley; 2010.
- Introductory materials can also be found at: http://www.john-uebersax.com/stat/biblio.htm

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Paul LaBrec, MS
Director, Research, 3M HIS
plabrec@mmm.com