OJPHI

Population Segmentation Using a Novel Socio-Demographic Dataset

Elisabeth L. Scheufele, MD, MS[1]; Brandi Hodor, BS[2]; George Popa, Jr., MS, MHSA[3]; Suwei Wang, PhD[3]; William J. Kassler, MD, MPH[4]

[1] Boston Children's Hospital, Boston, MA

[2] Merative, Ann Arbor, MI

[3] IBM Watson Health, Cambridge, MA

[4] Palantir Technologies, Denver, CO

Abstract

Appending market segmentation data to a national healthcare knowledge, attitude and behavior survey and medical claims by geocode can provide valuable insight for providers, payers and public health entities to better understand populations at a hyperlocal level and develop cohort-specific strategies for health improvement. A prolonged use case investigates population factors, including social determinants of health, in depression and develops cohort-level management strategies, utilizing market segmentation and survey data. Survey response scores for each segment were normalized against the average national score and appended to claims data to identify at-risk segment whose scores were compared with three socio-demographically comparable but not at-risk segments via Nonparametric Mann-Whitney U test to identify specific risk factors for intervention. The marketing segment, New Melting Point (NMP), was identified as at-risk. The median scores of three comparable segments differed from NMP in "Inability to Pay For Basic Needs" (121% vs 123%), "Lack of Transportation" (112% vs 153%), "Utilities Threatened" (103% vs 239%), "Delay Visiting MD" (67% vs 181%), "Delay/Not Fill Prescription" (117% vs 182%), "Depressed: All/Most Time" (127% vs 150%), and "Internet: Virtual Visit" (55% vs 130%) (all with $p<0.001$). The appended dataset illustrates NMP as having many stressors (e.g., difficult social situations, delaying seeking medical care). Strategies to improve depression management in NMP could employ virtual visits, or pharmacy incentives. Insights gleaned from appending market segmentation and healthcare utilization survey data can fill in knowledge gaps from claims-based data and provide practical and actionable insights for use by providers, payers and public health entities.

Keywords: Public Health, Social Marketing, Health Care Survey, Marketing Segmentation, Social Determinants of Health.

Abbreviation: designated marketing area (DMA), random digit dialing (RDD), index of concentration (IoC)

DOI: 10.5210/ojphi.v14i1.11651

Correspondence: Brandi Hodor, Merative, 100 Phoenix Drive, Ann Arbor, MI, 48108, 517-281-9336
bhodor@merative.com

Introduction

For decades, public health and healthcare leaders have advocated for the use of population segmentation and other techniques borrowed from the practice of marketing to improve delivery of services through more targeted approaches.[1,2] In the commercial sector, marketing uses segmentation to design and target products and services to meet the specific needs and desires of a particular group of consumers. However, to improve the effectiveness of programs and messages, social marketing in the health sector would involve the use of segmentation to more precisely tailor interventions and outreach to specific sub-populations.

Marketing has achieved success in changing consumer behavior by using segmentation models that include: knowledge, attitudes and beliefs; past behaviors; social norms and culture; and psychological characteristics such as needs, wants, values and lifestyle, readiness to change and future intentions. While targeting interventions based on traditional demographic characteristics (e.g., age, race, gender, educational level, income) reflects the basic tenants of epidemiology, social marketing proposes to refine that practice using a number of more personalized characteristics related to health in general to the specific outcome or behavior being sought. However, in spite of its potential, population segmentation remains underutilized in public health and healthcare, in part because of the lack of pertinent data and analytic approaches.[3]

This paper describes a novel approach for population segmentation, based on appending the geocoded responses from a national survey on health-related knowledge, attitudes and beliefs, with commercial market segmentation data containing sociodemographic data at the block group level. We further illustrate the additional benefits of linking these survey and sociodemographic data to administrative claims containing individual-level information on healthcare process and outcomes with a prolonged use case on the public health concern of depression.

In the US, people report suffering from depression at about a rate of 7-8%.[4,5] Depression also affects the workplace, as 27% of those affected report trouble with work or home.[4] The World Health Organization reports that in the US, every dollar used to treat depression results in 4 dollars of positive impact in productivity and health.[6] Businesses should invest in supporting the mental health of their employees as good mental health is also associated with optimizing employee work-related output and their overall well-being. The use case will address the issue of depression, both a major public health and workplace concern, by using the insights from PULSE® and PRIZM® to support better characterization of the population of concern, uncover barriers and opportunities for care, and organize outreach and management strategies to bridge gaps and overcome barriers to better mental health management.

Methods

The Survey

The PULSE® Healthcare Survey is a nationwide household survey reflective of health-related behaviors, attitudes and health care utilization habits of people in the United States, conducted annually since 1988. The Survey has two components: Core Topics and Survey Topics. Core Topics remain standard from year to year and include Health Status (e.g., Self-Perceived Health Status), Insurance Coverage, and Demographics (e.g., Respondent Age, Respondent Gender,

Household Composition). Survey Topics and related questions are changed annually to reflect current topics of interest to a variety of stakeholders. In total, there are about 100 Survey questions spread across the numerous topics. For instance, the 2019 PULSE Survey Topics included: Primary Care Utilization and Access, Social Determinants of Health, Telemedicine, Attitudes and Factors Influencing Physician Selection, Mental Health Status, and Avoiding Healthcare.

The PULSE Survey starts each year in January with the Pretest period and then is followed by 11 subsequent monthly waves. During each wave, a subset of the topics is surveyed. The Pretest period serves to determine: the survey length, the response rate, and the comprehension of the question by the respondent. All the questions are tested on 2000 participants during the Pretest. The information gleaned from the Pretest informs the selection of a subset of survey questions, how they are distributed over the 11 waves, and organized so the Survey can be executed in 11 minutes on average. Auditing is performed to ensure a threshold of 2000 (affirmative) responses has been achieved for each question to provide an appropriate level of reliability in survey output. To attain the primary goal of 80,000 respondents, 10 waves (March through December) have a threshold of 7250 participants and February has a threshold of 7500 which includes the January responses.

The Survey is multi-modal, and performed via landline telephone, with the addition of cell phones and internet modalities in 2013. The distribution of communication methods is 50% via phone (with 90% via landline, and 10% via cell), and 50% via the internet with sample matching, which is an internet survey method to allow incorporation of Survey answers from engaged rather than inattentive on-line participants. To ensure that distribution across the country is proportional to the percentage of households by the designated marketing area (DMA), random digit dialing (RDD) is employed which makes sure that the area code and the exchange portion of the number are tracked to ensure coverage, but the last four digits are randomized to obtain a location-controlled randomized sample. The RDD process results in about 60% of the Survey respondents, and participants from previous PULSE surveys make up the remaining 40%. The fielding is performed by a third-party vendor, who performs the interview portion of the survey exercise and delivers the Survey dataset.

The Market Segmentation System

The Claritas PRIZM® Premier Segmentation System is a market and consumer segmentation model that has been used by commercial entities to enhance their marketing and messaging strategies since the mid-1970s. PRIZM® Premier provides geo-demographic data by combining insights from data on consumerism behaviors with geographic data, down to the household level.

The market segmentation system is developed from several data sources that include the US Census and the supplementary census data, provided via "The American Community Survey." The data undergoes proprietary regression modeling called "Multivariate Divisive Partitioning," which is a modified Classification and Regression Tree method that allows partitioning to go across numerous elements to result in distinct behaviorally similar clusters.[7] The resultant 68 distinct nodes become the market segments, which are given descriptive names for more facile reference (e.g., "New Melting Pot", "Generation Web"). Claritas also employs a modeling process to establish Urbanicity Classes by defining the concept of urbanicity to improve the distinction

among geographic regions that may have similar population density but differentiate by urban vs rural type of experience.

The current version of 68 PRIZM Premier market segments are divided into 11 Lifestage Groups and 14 Social Groups, of which then roll up into 3 Lifestage Classes and 4 Urbanicity Classes, respectively (see Figure 1). The Social Groups are based on urbanicity and affluence, where the latter is defined by income, education and home value, with examples including U1 Urban Uptown or U2 Midtown Mix. These Social Groups roll up into 4 Urbanicity Classes, including Urban, Suburban, Second City, and Town & Rural. Two examples of market segments that roll up into Social Groups are Generation Web in City Centers and New Melting Pot in Micro-city Mix, all of which roll up to the Second City Urbanicity Class.[7] (These Social Groups and market segments will be revisited in the use case below).

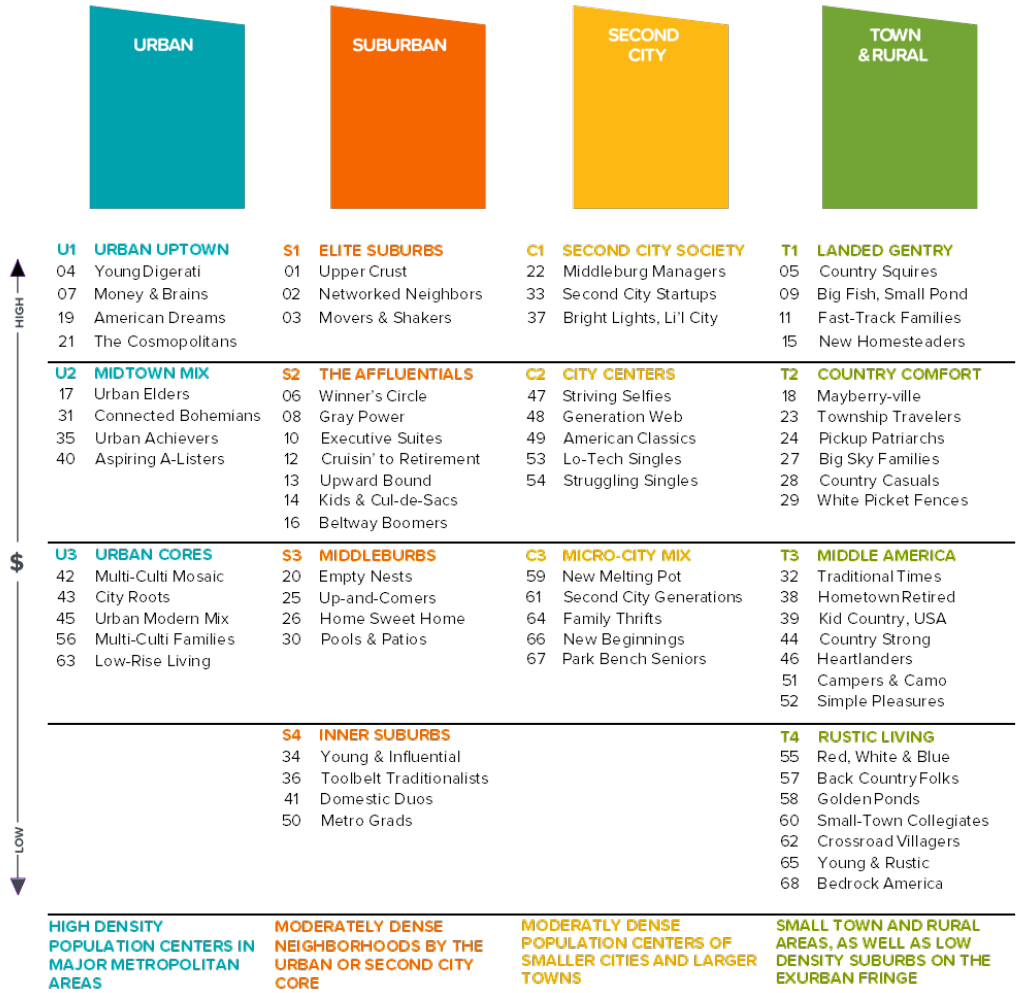

Figure 1: Claritas Market Segments Distributed by Urbanicity and Affluence. Reprinted with permission from Claritas®, Cincinnati, Ohio, United States.

Appending Marketing Segmentation to Survey Results

Appending the Segmentation to the Survey involves post-survey processing that includes weighting, appending by block group, and tabulating response rates. A weighting protocol is applied by the third-party vendor to the demographic data and to the responses by household and by individuals, to represent national demographics. As the Survey responses include the respondent address, the Segmentation data are then appended to the Survey responses by geocode. An age-cluster matrix of rates of the geocoded responses for each question are then tabulated for each of the more than 217,000 US block groups. These calculations are further broken down by age of Head of Household who answered the Survey questions, to align with the census data [age ranges of < 35, 35-64, >=65] to further calculate the response score by age group and block group.

The Index of Concentration (IoC) is calculated once the market segments are appended to the Survey responses. IoC is the ratio of the proportional responses by market segment clusters against the proportional responses of the national average. For example, consider that the national affirmative response rate to a question on whether the participant smokes is 20%. If the response rate for cluster A aggregates to 15%, and the response rate to cluster B aggregates to 25%, then the IoC for cluster A would be 15%/20%, or 75, and for cluster B would be 25%/20%, or 125. This would be interpreted as Cluster A is 25% less likely to smoke than the national average, and cluster B is 25% more likely to smoke than the national average. The IoC scores are centered to zero instead of 100%. Thus, if the IoC score is 125, the discussed score will be reported as “25% above the national average.”

Analysis

Claims data from an employer dataset with 19,616 members from a Midwest location were appended to the PULSE and PRIZM data at the block group level. An at-risk market segment was identified for study as the one with the highest affirmative score for PULSE response for depression and low claims utilization for depression. Then three segments in the same Urbanicity Class were identified, which would indicate some similarity in social demographic details, who have high utilization by claims for the diagnosis of depression. The at-risk segment was compared to the three comparable segments to identify potential reasons for the higher number of medical encounters on depression by the three comparable segments, implying better medical management for chronic depression. The Mann-Whitney-U test was applied to the original PULSE Survey IoC scores of the at-risk segment against the median scores of comparable segments to identify PULSE responses that are statistically significant in their differences. Then the appended Survey and Market Segmentation data was reviewed to better characterize the circumstances of each segment and identify specific barriers to healthcare which can be used to direct the design and targeting of interventions.

Human Participant Compliance

This study was reviewed and granted an exemption determination by the Western Institutional Review Board.

Results

Use Case: Characterizing an At-Risk Population, and Developing Public Health Strategy from Appending Claims Data to the PULSE/PRIZM Dataset

Second City, one of the Urbanicity groups, was initially explored and one of its market segments, New Melting Pot (with 334 members), was found to have the highest affirmative survey response for "Being Depression Most of the Time" (49% higher than the national average). This market segment was compared with three others who had the highest health care utilizations by encounters for the diagnosis of depression in the Social Group. Selected data on behaviors and affluence from the PRIZM market and selected responses from the PULSE survey as they apply to the use case of addressing depression were further investigated.

The claims for all members demonstrated an average of 4.9 members per 100 with a diagnosis of depression in a 12-month period (July 1, 2018 to June 30, 2019). Rates of depression were over 30% higher than the average for American Classics (6.7 members per 100 with encounters for depression) and Bright Lights, Li'l City (6.5 members per 100). The extremes were observed with Generation Web, who had a rate of 8.3 members per 100 with encounters for depression, which was 68% higher than the average, and New Melting Pot, who had a rate of 4.8 which was 2% less than the average, indicating that New Melting Pot had the least members per 100 with claims for the diagnosis of depression of these four segments.

The contrast in having the highest Survey response in reporting depression but with the lowest utilization rate via claims data could indicate that the New Melting Pot segment may have some unmet needs when it comes to mental health management. Low healthcare utilization can also suggest barriers to access. When the same populations were analyzed with the PULSE and PRIZM data, additional information can provide context to the situation in further insight into the subpopulations. In particular, the PULSE responses collected for this use case included those that provide value to characterizing the cohort at risk. These responses were evaluated via Nonparametric Mann-Whitney U test to detect significant differences between New Melting Pot vs the other three segments which provide insight into why they had higher reports of depression but less utilization than comparable segments and how to bridge those gaps (see Table 1).

Appending market segments to survey responses provided additional location-based perspectives (see Figure 2). In the analysis for depression, a focused subset of information from the PULSE survey brought to light some of the possible reasons why New Melting Pot would have a higher score for being "Depressed Most of the Time" when compared to the other 3 segments. New Melting Pot was associated with reports of numerous concerns, including "Worrying about Food" (123% higher than the national average) and "Utilities being Threated to be Turned Off" (138% higher), but also the "Safety of Their Neighborhood" (93% higher), all of which could impact or exacerbate their symptoms of depression while at the same time could make it difficult to focus on their healthcare. Understanding the types of drivers of stress may also explain disparities at the neighborhood level, and can allow public health, employer or health plan organizations to better understand and characterize these targeted populations and the households they were living in.

Table 1: Nonparametric Mann-Whitney U Test Comparing Median of Three Comparable Segments Against Segment at Risk. The test performed on the original IoC values for the PULSE responses (p-value compares the New Melting Pot segment with the Median value of the other three segments - Bright Lights, Li'l City, Generation Web, and American Classics).

Cluster	Median	New Melting Pot	p-value
Ability to Pay for Basic Needs - Somewhat Hard	121	125	< 0.001
Lack of Transportation Affected Daily Living	112	153	< 0.001
Utilities Threatened to Turn Off - Yes	103	239	< 0.001
Delay Visiting MD Next 3 Months - Highly Likely	67	181	< 0.001
Delay/Not Fill A Prescription Next 3 Months - Highly Likely	117	182	< 0.001
Depressed: All/Most Time	127	150	< 0.001
Internet: Using App/Software to Manage Health	106	123	< 0.001
Internet: Had Virtual Visit Past 12 Months	55	130	< 0.001
Neighborhood Safe for Family - Strongly Disagree/Disagree	61	193	< 0.001
Struggle Put Food on Table - Strongly Agree/Agree	81	102	< 0.001

Continuing with the issue of depression, the Survey data could help identify barriers to accessing healthcare while the socio-demographic data and health perspective of the Market Segmentation data provided further insight into the gaps in care. New Melting Pot reported higher levels of "Delay Seeing the Doctor" (79% higher than the national average), reported more frequently "Delaying Filling Their Prescriptions" (81% higher), and reported higher concerns with "Lacking Transportation" (52% higher) (see Figure 2). From the Segmentation data, people in New Melting Pot were associated more often with a high school reading level and with working primarily in the services industry. These insights indicated lack of basic resources for this market segment, such as consistent access to transportation, work schedules that may not be flexible enough to accommodate weekday visits to the doctor, poor access to affordable medications or prescription plans, and potential challenges in health literacy. The Survey data can help identify specific factors and drivers not typically available in claims data and can provide additional perspective to help focus on the needs of a cohort at risk.

The PULSE/PRIZM dataset can be used to develop strategies for addressing depression management for the at-risk population. As discussed above, New Melting Pot was a resource low market segment, struggling to manage basic needs including food and utilities, and lacking in transportation options. They also had trouble filling their prescriptions and making it to their doctor's appointment. However, information from the Survey data suggested there may be some options to fill the gap in care. New Melting Pot reported above average "Use of Virtual Visits in the Last 12 Months" (28% higher than average), and higher reports of "Having Used/Using an App for Health Care" (20% higher) (see Figure 2). Strategies to enable people in New Melting Pot

might include providing travel vouchers, such as for ride shares, to help with transportation. Another option can be to support and provide telehealth access, as this group had reported higher use than average and also was technically inclined. Education should be provided at the appropriate level to facilitate comprehension. Prescription vouchers programs could be directed to these patients for any medication needs to help with their mental healthcare. Organizations can use these insights to better strategize for campaigns to target members that will resonate with their needs and capabilities, moving beyond a blanket approach for healthcare engagement and applying more personalization at the community level.

Figure 2: Four Market Segments Identified to Evaluate Depression in Cohorts. The claims dataset has been appended to PRIZM Market Segmentation and to PULSE Survey responses, with use case pertinent information from each dataset detailed by market segment, and also PULSE responses centered to zero for ease of discussion.

Discussion

This paper presents a unique approach to aggregating information that is often not available in clinical or claims datasets. Data from a US annual national healthcare survey appended to a market segmentation model can be further combined with patient-level data by geocode to generate very precise population insights at a hyperlocal level. Just as commercial marketing aims to use population segmentation techniques to impact consumer behavior, healthcare and public health can use these techniques to impact health-related behavior. Segmentation into precision cohorts can identify groups with specific needs and risks, and furthermore helps to target specific communication messages and outreach strategies that match the targeted community. Social marketing applied to health concerns is hard to implement, and these data and analytic methods may provide practical resources to operationalize this approach.

Information from the PULSE/PRIZM appended dataset can bring sociodemographic details and attitudes and behavior on health and health care of patients at the community level that can help better elucidate the perspectives and barriers of a group of patients. For the illustrated use case of depression, the PULSE survey data provided possible reasons for why the group may be depressed, demonstrating that the cohort may lack resources, socioeconomic status and access to care. The PRIZM marketing data demonstrated that the group is under-resourced with specific barriers to accessing care. Knowing that this cohort is amenable to virtual health visits and understanding that this group tends to delay or not see a doctor, suggested the opportunity to overcome these barriers through telehealth visits.

Applying the PULSE/PRIZM dataset to better understand a patient population is, at its core, an ecological analysis of a cohort of people. Consider that the strength of such an analysis is strongest at the population level, where it provides broad knowledge localized to a cohort of people. However, a limitation includes when insights are considered at the individual level. Because people tend to self-assort with those who are similar, survey and marketing characteristics are highly representative of the predominant group at the block level. Although peers and neighbors may generally have similar or overlapping tendencies and behaviors, any specific individual may not be representative of those generalizations. One of the means by which to address this limitation is the addition of appended patient level claims, which can provide the specificity that is not available from the PULSE/PRIZM dataset alone. The appended super dataset speaks probabilistically to a block group, but there may be individuals in the block group for which this does not apply. Thus, caution should be used when generalizing to the individual in the group, which can be mitigated with individual level claims data.

Conclusion

Enriching claims data with hyperlocal insight from a market segmentation model and a national health care survey can fill gaps in socio-demographic, behavioral, and attitudinal information not normally found in clinical datasets. Public health agencies and healthcare organizations that seek to optimize the health of a population under their purview (e.g., insurers, public health entities, employers) require more information than what is present in commonly available datasets to better understand their patients' social needs and develop community-level focused strategies on issues of concern. Insights from PULSE and PRIZM can be applied to claims data to better characterize

a neighborhood or population group at risk to better understand their vulnerability, identify barriers to management or care, and more precisely target outreach.

Acknowledgements

Not applicable

Financial Disclosure

No funds, grants, or other support was received.

Competing interests

All authors were employees of IBM Watson Health at the time this study was conducted.

References

1. Slater MD, Flora JA. 1991. Health lifestyles: audience segmentation analysis for public health interventions. *Health Educ Q*. 18(2), 221-33. doi:https://doi.org/10.1177/109019819101800207. PubMed

2. Lynn J, Straube BM, Bell KM, Jencks SF, Kambic RT. 2007. Using population segmentation to provide better health care for all: the "Bridges to Health" model. *Milbank Q*. 85(2), 185-208. doi:https://doi.org/10.1111/j.1468-0009.2007.00483.x. PubMed

3. Vuik SI, Mayer EK, Darzi A. 2016. Patient Segmentation Analysis Offers Significant Benefits For Integrated Care And Support. *Health Aff (Millwood)*. 35(5), 769-75. doi:https://doi.org/10.1377/hlthaff.2015.1311. PubMed

4. Prevalence of Depression Among Adults Aged 20 and Over: United States, 2013–2016. 2018 [Accessed January 19, 2021]; Available from: https://www.cdc.gov/nchs/products/databriefs/db303.htm

5. Major Depression. 2019 [Accessed January 19, 2021]; Available from: https://www.nimh.nih.gov/health/statistics/major-depression.shtml

6. Mental health in the workplace. 2020 [Accessed January 19, 2021]; Available from: https://www.who.int/mental_health/in_the_workplace/en/

7. Claritas PRIZM Premier Methodology 2019, copyright 2018