Prioritizing Which Homeless People Get Housing Using Predictive Algorithms

An Evidence-Based Approach to Prioritizing High-Cost and High-Need Homeless Persons for Permanent Supportive Housing

Halil Toros and Daniel Flaming

ECONOMIC ROUNDTABLE
Knowledge for the Greater Good

www.economicrt.org

Prioritizing Which Homeless People Get Housing Using Predictive Algorithms

An Evidence-Based Approach to Prioritizing High-Cost and High-Need Homeless Persons for Permanent Supportive Housing

Halil Toros and Daniel Flaming

We present a predictive model for identifying homeless persons likely to have high future costs for public services. It was developed by linking administrative records from 2007 through 2012 for seven Santa Clara County agencies and identifying 38 demographic, clinical and service utilization variables with the greatest predictive value. 57,259 records from 2007 to 2009 were modelled, and the algorithm was validated using 2010 and 2011 records to predict high cost status in 2012. The model generated a good area under the ROC curve of 0.83. A business case scenario shows that two-thirds of the top 1,000 high-cost users predicted by the model are true positives with estimated post-housing cost reductions of over $19,000 per person in 2011. The model performed very well in giving low scores to homeless persons with one-time cost spikes, achieving the desired result of excluding cases with single-year rather than ongoing high costs.

Keywords: homelessness; permanent supportive housing; homelessness prevention; triage tool; predictive analytics; public costs

Overview

Homelessness is a major social problem in the United States, with large public health impacts affecting millions of individuals and families and costing billions of dollars. The most recent numbers available from the Department of Housing and Urban Development (HUD) show almost 1.5 million persons experiencing homelessness at some point over the course of a year, with the number increasing by approximately five percent from 2013 to 2014. The national rate of homelessness was 47.2 homeless people per 10,000 people in the general population. This estimate does not include people in unsheltered locations who never accessed a shelter program during the year (HUD, 2014).

Growth in homelessness over the last three decades has been exacerbated by economic downturns, loss of affordable housing and foreclosures, stagnating wages, an inadequate safety net, and the closing of state psychiatric institutions. The federal response to homelessness changed in 2009, with the creation of the federal Homelessness Prevention and Rapid Re-housing Program (HPRP). The purpose of HPRP was to provide homelessness prevention assistance to households that would otherwise become homeless—many due to the economic crisis—and to provide assistance to rapidly rehouse persons who were homeless (HUD, 2009). This new program represented a paradigmatic shift from uncoordinated short-term responses to avert homelessness, primarily using shelters, to prioritizing homelessness prevention.

The first component of this strategy is expanding permanent supportive housing opportunities for people experiencing chronic homelessness, and prioritizing those with the most severe challenges for assistance. The second component is connecting permanent supportive housing to street outreach, shelter, and institutional "in-reach" to identify and engage people experiencing chronic homelessness. The third component is community-wide adoption of *Housing First*, that is providing permanent housing as quickly as possible, and increasing federal funding to expand the national inventory of permanent supportive housing by 25,000 units in order to end chronic homelessness and prevent its recurrence (USICH, 2015). The scaled-up federal funding needed to end chronic homelessness still awaits approval by Congress.

The housing first model was introduced by the New York City based nonprofit organization, Pathways to Housing, to provide homelessness intervention services to adults with psychiatric diagnoses and substance abuse problems. They provided immediate housing and services to homeless adults with co-occurring diagnosis as a matter of right, with no pre-conditions. They also incorporated a harm reduction approach to psychiatric and substance abuse treatment, and empowered the consumers of services to make choices about housing and services (Greenwood, Stefancic and Tsemberis, 2013).

The proposed new initiative prioritizes housing stabilization as the centerpiece of homelessness assistance. A central theme is that some people need more than housing assistance to stabilize. A small but highly visible segment of the chronically homeless population has substantial service needs. To serve these needs effectively, the new initiative prioritizes providing permanent supportive housing (PSH) using a housing first approach as the solution to chronic homelessness. PSH programs are designed to serve homeless individuals with disabilities that interfere with maintaining housing on their own by providing permanently affordable housing combined with ongoing supportive services to help them become stable renters.

For people experiencing chronic homelessness, research shows that PSH using a housing first approach is an effective intervention for enabling individuals to exit homelessness and for preventing its reoccurrence in the future. However, since housing resources are limited, one of the key challenges of the intervention is identifying and targeting the "highest priority" individuals so as to allocate this scarce resource in a way that produces the greatest benefit. It is well documented that costly interventions, such as PSH, are not likely to generate cost offsets equal or higher than the cost of the interventions, except for the most costly users (Culhane, 2008; Paulin et al., 2010).

This paper presents a triage tool to improve the efficiency of PSH by identifying high-cost homeless persons for whom the solution of housing costs less than the problem of homelessness. It is expected that the triage tool will achieve more efficient allocation of housing resources, creating cost offsets from reduced service use after people are stably housed that can be stretched across a larger pool of homeless people whose housing can be subsidized with those offsets.

The purpose of the triage tool is to identify persons experiencing homelessness with a history of costly utilization of public services so that they can be connected permanently affordable housing and cost-effective community-based health care and support services. The tool applies a statistical predictive model to administrative data in order to prioritize homeless adults with the highest needs and public costs. The intent is to provide a highly accurate predictive model comparable to those developed through studies of high-cost health system users. These models were developed to identify patients at high risk of readmission to a hospital based on demographics, prior hospital admissions and clinical conditions (Ash, et al. 2001; Billings, et al. 2006, 2013; Chechulin, 2014; Fleishman and Cohen, 2010; Moturu, Johnson and Liu, 2010; Tamang et al., 2015).

The Silicon Valley Triage Tool improves on earlier tools developed by Economic Roundtable to identify the one-tenth of homeless individuals with the highest public costs, and the acute ongoing crises that create those high costs (Economic Roundtable, 2011, 2012). In addition to the work done by Economic Roundtable in identifying high-cost homeless persons, a few other studies have used predictive models to assess homeless risks. Byrne et al. (2016) estimated predictors of homelessness and developed methods for more efficiently targeting homelessness prevention services. A recent study on New York City Home Base prevention program for families showed that adoption of an empirical model for deciding which families to serve can make homelessness prevention more efficient (Shinn et al., 2013). And the Veterans' Administration has been working on using predictive models in screening homeless Veterans (Montgomery et al., 2013).

This paper extends previous research applying predictive models to homelessness and high-cost service users. The model presented in this paper predicts who will or will not become a high-cost public service user in the next year, given various person-level characteristics in the current year and previous year, providing a predictive score (probability) for each individual in order to determine priorities for intervention across large numbers of individuals.

In this paper we describe the predictive modeling methodology used to develop a triage tool to prioritize housing access for an efficient and cost effective PSH program. After presenting the results and validation of the model, we develop a business scenario to estimate the cost savings after the implementation of the triage tool. The paper ends with a discussion of potential ways to utilize the tool in practice, limitations and recommendations.

Chronic Homelessness

The majority of people who become homeless remain so for less than a year. A smaller number of people, however, remain homeless much longer, experiencing continuous and chronic homelessness. According to federal guidelines, an individual is chronically homeless if he or she has a diagnosed disability—such as serious mental illness, substance use disorder, posttraumatic stress disorder, cognitive impairments or chronic physical illness or disability—and has been homeless and lives in a place not meant for human habitation, a safe haven, or in an emergency shelter for at least one continuous year or has experienced at least four episodes of homelessness in the past three years where the cumulative total of the four occasions is at least one year (Federal Register, 2015).

The prevalence of chronic homelessness can be estimated using HUD's point-in-time (PIT) data. On a single night in January 2014, communities across the country counted 578,424 homeless persons, of whom 99,434 or 17 percent were chronically homeless. Among individual adults, 23 percent were chronically homeless. Among family members, seven percent were chronically homeless (HUD 2014). It was estimated earlier that, over the course of a year between 150,000 and 250,000 single adults experience chronic homelessness (Burt 2002). Several earlier studies have also shown prevalence rates between 10 percent and 27 percent depending on the definition used, the duration of time observed, and the method of data analysis (Burt et al., 2005; Caton et al., 2005; Kuhn and Culhane, 1998). Recent research in Santa Clara County found that 13 percent of individuals who were homelessness during a six-year interval experienced chronic homelessness (Economic Roundtable, 2015).

Needs of chronically homeless individuals that are essential for their well-being go unmet, including connections to housing, income, family, and health. This leads to stress, anxiety, depression, deprivation and chaos, destabilizing their lives. Over time, chronically homeless individuals have increasingly complex and costly needs, including serious health and mental health conditions and disabilities that result in cycling in and out of hospitals, jails, prisons, psychiatric hospitals, and homeless shelters.

Several studies describe the clinical and social characteristics and patterns of service utilization among people who are chronically homeless. The majority of individuals have a serious mental illness such as schizophrenia, bipolar disorder, or major depression. They also experience high rates of substance abuse disorders, physical disability, or chronic disease. Many experience co-occurring mental illness and substance use problems (Burt, 2002; Caton et.al, 2005, 2007; Folsom et al., 2005; Rosenheck, 2000). In addition to serious disability, the lives of chronically homeless people are compromised by persistent unemployment and a lack of earned income forcing dependence on public assistance for sustenance, health care, and, if fortunate, an eventual exit from homelessness (Caton et al., 2005, 2007). Moreover, chronically homeless individuals often have a long arrest history, cycling through jail and prison (Caton et al., 2005; Kushel et al., 2005; Metraux and Culhane, 2004; Zugazaga, 2004).

The health, personal, and economic challenges that chronically homeless individuals experience and the lack of effective, coordinated services to address these problems often lead to a vicious cycle of diminished well-being with serious implications for their service utilization patterns. Their impairments impede access to needed health services and other support systems such as employment services. Consequently, they cycle through costly emergency-driven public systems without getting the ongoing care they need to address severe mental illness, substance use disorders, or chronic health conditions (Caton et al., 2007; Folsom et al., 2005).

Chronically homeless individuals spend a disproportionate number of days in the shelter system (Kuhn and Culhane, 1998; Metraux et al., 2001). In addition, because of their complex and co-occurring disabling conditions, poor health status and elevated rates of unintentional injuries and traumatic injuries from assault, chronically homeless persons have high rates of hospital emergency rooms use and hospitalization, and longer hospital stays for mental health and substance abuse problems (Culhane, Metraux and Hadley, 2002; Folsom et al., 2005; Kuno et al., 2000; Kushel et al., 2002). As the chronically homeless population ages, their utilization of emergency rooms and hospital rooms increase (Caton, et. al 2007). High incarceration rates coupled with heavy use of mental health and medical facilities in jails and prisons are also well documented (Kushel. et al., 2005; McNiel, Binder and Robinson, 2005; Metraux and Culhane, 2004).

Heavy use of acute and behavioral health care, criminal justice involvement, and use of social services costs tens of thousands of dollars per individual annually (Culhane, Metraux and Hadley, 2002; Martinez and Burt, 2006; Gilmer et al., 2009; Larimer et al., 2009; Bcom and Larimer, 2015; McLaughlin, 2011). While chronically homeless people represent only 20 percent of shelter users, they consume the largest share of health, social, and justice services with enormous costs (Ly and Latimer, 2015). In Los Angeles County, among homeless General Relief program participants, studies showed that the highest cost decile accounted for 56 percent of all public costs for homeless single adults (Economic Roundtable, 2009, 2011). A recent study using Santa Clara County data also showed that homeless costs are heavily skewed toward a comparatively small number of frequent users of public and medical services. Among residents experiencing homelessness in 2012, the 10 percent with the highest costs, the tenth decile, accounted for 61 percent of all public costs for homelessness and the top five percent accounted for 47 percent of all costs (Economic Roundtable, 2015).

Federal funding for homeless programs increased from $3.7 billion in 2010 to nearly $5.5 billion in 2016 (USICH, 2016). In addition, there are federal expenditures for homeless individuals through Medicaid, Medicare and the Veteran Administration, as well as large expenditures by state and county governments and institutions such as hospitals, jails and social service agencies.

Even though there have been growing public outlays to address chronic homelessness since 2010, the prevalence and costs of homelessness remain high. With finite resources for homeless assistance, prevention services and cost-effective interventions such as permanent supportive housing have attracted growing interest from policymakers and academic research over the past decade (Apicello, 2010; Burt et al., 2005; Byrne, et al., 2014; Culhane, Metraux and Byrne, 2011).

Preventive Services and Permanent Supportive Housing

The logic of prevention requires the definition of what is to be prevented (such as chronic homelessness) and the specification of the services with an association (preferably causal) between the intervention and the prevention of the undesirable condition using a series of risk and protective factors. Several frameworks have been suggested for developing prevention strategies for homelessness (Burt et al., 2005). The high-risk framework is the most appropriate framework for conceptualizing how to design homelessness prevention policies because it draws attention to the need for direct intervention among those at greatest risk. This framework focuses on alleviating the causes of homelessness for the most vulnerable subpopulations (Apicello, 2010).

To be successful, prevention strategies for high-risk individuals need to be both effective and efficient (Burt et al., 2005; Culhane, Metraux and Byrne, 2011; Shinn, Baumohl and Hopper, 2001). In this context, effectiveness refers to how capable a program is of facilitating the desired goal - prevention of homelessness with reasonable costs. Effectiveness should be evaluated with robust designs by comparing a treatment group of persons who received services to a control group of individuals not subject to the intervention. Otherwise, the effect of the services in preventing homelessness cannot be assessed accurately, because it is unrealistic to assume that all of the people who received services would have become or stayed homeless in the absence of those services. It is also possible that the effect of services might have not been significant; homelessness might have been merely postponed; or the ranks of high-risk individuals might simply have been reshuffled, allowing some to "jump the queue" and push others back in the line (Shinn, Baumohl and Hopper, 2001).

As noted earlier, recent research has shown that PSH using a housing first approach is a very effective homeless prevention service and has led to widespread and successful efforts to reduce chronic homelessness (Byrne et al., 2014, Culhane, Metraux, and Hadley, 2002; Greenwood, Stefancic and Tsemberis, 2013; Larimer et al., 2009; Rog et al., 2014; Tsemberis and Eisenberg, 2000; USICH, 2010, 2015). Based on increasing evidence, the U.S. federal government has endorsed PSH using a housing first approach as the ‘‘clear solution’’ to chronic homelessness and the PSH has become an important priority for HUD. The number of beds in PSH projects increased by almost 60 percent between 2007 and 2014, when an estimated 285,400 people lived in PSH (HUD, 2014; USICH, 2010).

Research has also demonstrated the effectiveness of PSH in generating cost-offsets. Many studies have shown that PSH and housing first interventions for chronically homeless population lead to cost savings through reduced shelter costs, decrease in both psychiatric and medical inpatient hospitalization costs, lower emergency room visit costs, reduced substance abuse treatment costs, and reduced criminal justice costs due to fewer arrests, detentions and court appearances (Henwood et al., 2015; Bcom and Latimer, 2015; Culhane and Byrne 2010;

Martinez and Burt, 2006; Shinn, Baumohl and Hopper, 2001, 2013; Toros and Stevens, 2012). Cost savings from providing PSH to homeless people with mental disorders was shown to be substantial (Culhane, Metraux, and Hadley, 2002; Gilmer et al., 2009; Larimer et al., 2009; McLaughlin, 2011; Sadowski et.al, 2009).

Despite such successes, the high cost of PSH would limit its availability to chronically homeless individuals with the greatest service needs if cost offsets are the benchmark for determining eligibility. Culhane (2008) reviewed several studies and concluded PSH is not likely to generate cost offsets equal to the cost of the interventions, except for the most costly users. Other studies also support the view that only frequent users of higher-cost services are likely to have sufficiently high costs to fully or mostly offset the costs of a PSH placement. Some research indicates that group may be limited to the most costly 10 percent of the chronically homeless (Poulin, et al., 2010; Roesenheck, 2000). Moreover, since homeless people are typically placed in PSH programs at times when they are in crisis and have had relatively high service use, regression to the mean results in decreasing costs for many of these people, even if they are not placed in PSH (Bcom and Latimer, 2015).

Hence, the research demonstrates that while PSH is effective in reducing chronic homelessness and yields significant cost offsets, to be efficient, it should target high-cost homeless persons so that off-sets will cover program and housing costs. In the context of homelessness prevention, efficiency refers to targeting high-risk individuals. Efficient targeting is critical in the design and success of prevention services (Apicello, 2010; Burt et al., 2005; Culhane, Metraux and Byrne, 2011; Shinn et al., 2001). An efficient program should use empirically and/or theoretically derived risk factors to identify high-risk individuals who are likely to stay homeless and use costly public services unless they receive the prevention services.

However, the efficiency criterion introduces a serious challenge. Predictive models and screening tools are subject to the well-known trade-off between sensitivity (the probability of correctly identifying true positives or those who will stay high-cost homeless persons in the absence of the prevention program) and specificity (the probability of correctly identifying true negatives or those would stay as low-cost homeless persons). If a low cutoff is selected, while the sensitivity increases and the model capturing more true positives, the specificity decreases leading to higher numbers of false positives. On the other hand, if the targeting cutoff is increased there are fewer false positives but many true positives are missed. This difficult trade-off is at the core of the efficiency issue, as savings realized through placing a high-cost homeless person in PSH will be washed out if many low-cost homeless persons are also placed (Culhane, 2011).

In the literature it is argued that the common failing of many prevention efforts is their targeting inefficiency, which leads to ineffective programs (Burt et al., 2005). It is also argued in the literature that available screening models are not sensitive or accurate enough to yield high hit rates without missing a large number of high-risk persons who would benefit from the program while producing cost savings (Apicello, 2010; Shinn, Baumohl and Hopper, 2001). However, recent technological advances in the fields of predictive analytics and data mining together with the availability of digital integrated administrative datasets with rich service utilization fields allow significant improvement in prediction ability over earlier approaches and models (Larson, 2013).

This paper presents the Silicon Valley Triage Tool for identifying homeless individuals in jails, hospitals and clinics who have continuing crises in their lives that create very high public costs. The model is very robust and accurate, taking advantage of advanced prediction methodologies and a unique and exceptionally valuable database created by Santa Clara County, home to Silicon Valley, linking service and cost records across county departments for the entire population of residents who experienced homelessness over a six-year period – a total of 104,206 individuals. The tool accurately identifies individuals experiencing homelessness whose acute needs create the greatest public costs and is expected to serve as a screening tool for efficient and effective PSH programs.

Methods

Data

By collaborating in linking their client records, seven agencies (HUD Continuum of Care Board, Criminal Justice Information Control system of the Sheriff Department, Department of Alcohol and Drug Services, Emergency Management System, Mental Health Department, Social Services Agency, Valley Medical Center) in Santa Clara County provided information on medical care (in-patient and out-patient), Emergency Medical services (EMS), and ambulatory care, drug and alcohol treatment services, mental health treatment services (in-patient and out-patient), incarceration (arrest, court and medical and mental health services in custody), and HUD funded social and homelessness services (Economic Roundtable, 2015).

In some instances, such as with demographic information (age, gender and ethnicity) and medical diagnoses, the same information is aggregated from multiple agencies to ensure that it is complete. A population of 57,259 homeless persons was used to develop the tool. These were individuals with at least one record linked to an agency during our six-year study window from 2007 through 2012. Due to record linkage problems experienced across some agencies, only those individuals with a homeless service use during the study window who also had a record in any of the other agencies contributing data were included in the cohort.

The model predicts the high cost status (defined as being in the top 10 percent of the homeless persons with highest public services costs) in 2009, using person characteristics from 2007 and 2008. Most of the cost components were included in the data with the exception of homelessness service costs provided by HUD Continuum of Care Board and criminal justice costs, which were derived using cost factors.

Data from 2007 to 2009 comprised the training sample. The validation was conducted by applying the model to 2010 and 2011 records to predict high cost status in 2012. The sample size for the training and validation cohorts was 57,259 records. The target group was 5,726 homeless individuals who made up top 10 percent with the highest costs. The validation cohort (2010-2012) was necessary to assess the out-of-sample predictive power of the model. Strong predictive power is often observed based on in-sample performance if the model over-fits the data. When that is the case, cases the model only explains well the training data, and out-of-sample performance is very poor. Since a predictive model is intended to be applied to new data with unknown outcomes, validation is needed to assess a model's performance.

Measures

Linked datasets provided information about factors that affect the outcome of interest - being a high cost user next year. These included demographic variables (e.g., age, gender, ethnicity); clinical variables (e.g., ICD-9 medical diagnoses), and utilization variables for all service types from the current and previous year (e.g., number of clinic or emergency room visits, number of hospitalizations, number of arrests), as well as the cost of services.

The binary target variable indicated whether or not homeless persons were top 10 percent of high-cost users in 2009 (training cohort) and 2012 (validation cohort). In order to identify high cost status, costs were summed across all service types and then ranked separately for the training and validation cohorts.

Model development was conducted in two stages. In the pre-processing stage, potential variables that would have an effect on becoming a high-cost user were identified based on earlier research and a series of F-tests (for categorical variables) and t-tests (for continuous variables). This step generated the first iteration of variable selection after eliminating redundant and irrelevant factors with p-values greater than 0.25. The initial set of selected variables was transformed and prepared for model development using several techniques such as binning continuous variables, clustering categorical variables, and generating binary and count variables. All variables were generated for the current and previous years and a total of 256 input variables were selected to be included in the model development.

Analysis

Several models for predicting high-cost users were developed and their performance was assessed using the SAS Enterprise Miner platform (Sarma, 2013; SAS, 2013). Several regression techniques were implemented to build models predicting the status of each person in the dataset as a high cost user in the next year. Since the predictive model is intended as a triage screening tool, it must utilize factors that contribute to the most accurate final score or probability possible, and assign weights proportionate to each factor's effect. Knowing the input factors used in the model is critical in building the logistics of data integration behind this model. Hence, we tested three techniques; logistic regression, least-angle regression and decision tree models that are capable of explaining the classification or decision process rather than using machine-learning algorithms that do not explain how given types of information are used to make predictions.

A comparison of the models' performance based on the receiver operating characteristics (ROC) curve led to selecting a logistic regression model as the champion model. In the final phase this model was fine-tuned, introducing interactions between variables, testing the non-linearity of variables and applying a sensitivity analysis to decrease the number of variables - particularly testing if current and previous year variables could be aggregated into a single variable without sacrificing the model's performance.

The final model was validated using the 2010-2012 cohort to assess the out-of-sample predictive power of the model. Sensitivity, specificity, positive predictive value (PPV), and accuracy measures as well as the area ROC curve were used to assess the out-of-sample model performance (See Gonen, 2007).

The sensitivity statistic measures the proportion of high-cost homeless persons correctly identified by the model with high scores. It is also known as the true positive rate and reflects how well the model performs in capturing those homeless persons with high future costs. If the level is too low, a large number of high-cost homeless persons would not be provided with permanent supportive housing.

The specificity statistic measures the proportion of not-high-cost homeless persons correctly identified by the model with low scores. If the level is too low, this is translated into to a high false positive rate (1-specificity) meaning a large number of low cost homeless persons would be provided with permanent supportive housing.

The PPV statistic estimates the accuracy of the model by measuring the proportion of true positives (correctly classified high-cost homeless persons) within the population of all persons identified as high-cost persons. In other words, it is the probability that persons with a high score (above a defined cost threshold) truly are high-cost persons. Finally, the accuracy statistic measures the proportion of true positives and true negatives out of all persons.

The validated model was later utilized to estimate the potential costs and benefits of applying the model under several cut-off thresholds and making certain assumptions about costs of PSH and likely reduction in service use attributable to the PSH placement.

Results

The final model had 38 variables with main effects and 11 variables with interactions. The descriptive values of model variables are shown in Table 1. Significance of the parameter estimates (p-values) and odds ratios are presented in Table 2. As shown in Table 1, high-cost homeless persons in Santa Clara County represent a higher proportion of males than the overall population that experienced homelessness, and are slightly older. Their rate of engagement in the criminal justice system is very high relative to the rest of the population. Almost half of them were arrested during the previous two years compared to only 16 percent for the rest of the population. Their average number of days in jail is more than 6 times greater than the rest of the population - 32.9 days vs. 5.2 days.

After testing 970 3-digit ICD-9 medical diagnoses, 43 diagnostic groups, and 18 body system diagnostic categories, the model retained six effective diagnosis codes or groups—adjustment reaction, organ failures, heart diseases, schizophrenia, neoplasm, and other ill-defined and unknown causes of morbidity and mortality. In addition, two other factors were included, which are the aggregations of chronic medical conditions and high-cost ICD-9. The high-cost homeless group shows much higher rates of encounters with these diagnoses while overall averages vary between six percent (heart diseases) and 68 percent (chronic medical condition). More than half of the high-cost group had been diagnosed with one or more of the 59 high-cost ICD-9s, while only a fifth of the lower-cost population had any of these diagnoses.

The high-cost group also shows higher rates of engagement with health and emergency services. There are large group differences for emergency medical service encounters (30 percent vs. 7 percent), hospital inpatient admissions via emergency room admission or transfer from a psychiatric facility (20 percent vs. 4 percent) and outpatient psychiatric emergency services or ambulatory surgery (41 percent vs. 15 percent). The number of admissions and days of inpatient hospitalization, and number of outpatient encounters are also significantly higher for high-cost homeless persons.

Finally, behavioral health data show more frequent encounters for the high-cost group. Both mental health (inpatient and outpatient) and substance abuse service rates are higher. The prevalence of documented substance abuse, as indicated by any recorded medical diagnosis or justice system charge, is twice as high for the high-cost group – 61 percent vs. 31 percent for the balance of the population. In contrast, there is little difference in public assistance and homeless service participation rates.

Table 1. Averages of Model Variables for High Cost and Other Homeless Persons (Validation Sample)

Variable (all values are percentages or means)	High Cost (N=5,726)	Other (N=51,533)
Demographics		
Age less than 18	5%	10%
Age 18-45	56%	55%
Age 46-65	36%	31%
Age 65+	3%	4%
Female	42%	54%
Criminal Justice		
100+ days of probation in the last 2 years	18%	5%
Arrested in last 2 years	46%	16%
Jail booking in last 2 years	23%	9%
Jail security classification of 3 or 4 (i.e., high risk) this year	10%	1%
Arrested for inebriation and released within 48 hours - this year	8%	1%
Number of arrests this year	.78	.16
Number of days in jail this year	32.9	5.2
Health Diagnoses		
Diagnosed with chronic medical condition Chronic Condition Indicator for ICD-9-CM diagnosis codes by the Healthcare Cost and Utilization Project (HCUP)	68%	35%
Medical encounter with diagnosis of adjustment reaction ICD-9 309 in last 2 years	11%	3%
Medical encounter with diagnosis of heart disease ICD-9 401-429 in last 2 years	6%	2%
Number of medical encounters with diagnosis of organ failure ICD-9 569-573, 576-578, 585-594, or 596 in last 2 years	.6	.1
Medical encounter with diagnosis of schizophrenia ICD-9 295 in last 2 years	14%	2%
Number of medical encounters with diagnosis of neoplasm (ICD-9 140 to 239) in last 2 years	.4	.1
Medical encounter with diagnosis of "other ill-defined and unknown causes of morbidity and mortality" (ICD-9 799) in last 2 years	17%	4%
Medical encounter with diagnosis of high-cost ICD-9 in last 2 years	52%	20%
Health & Emergency Services		
Emergency Medical Service (EMS) encounter this year	30%	7%
Emergency Medical Service (EMS) encounter last year	29%	7%
Two or more Emergency Medical Service (EMS) encounters in last 2 years	12%	1%
Admitted as hospital inpatient via emergency unit admission or transfer from psychiatric facility in last 2 years	20%	4%
Outpatient Psychiatric Emergency Services or ambulatory surgery this year	41%	15%
Number of hospital inpatient admissions this year	.3	.06
Number of hospital inpatient days in last 2 years	3.7	.6
Non-inpatient (ER or clinic visit) health system encounter this year	68%	43%
Number of non-inpatient (ER or clinic visits) encounters this year	6.2	2.3
11+ non-inpatient (ER or clinic visits) health system encounters this year	20%	6%
Behavioral Health		
Number of Mental Health outpatient days in the last 2 years	11.1	2.1
Two or more Mental Health outpatient visits in the last 2 years	27%	9%

Number of Mental Health inpatient admissions this year	17.6	1.2
Two or more Mental Health inpatient admission in the last 2 years	20%	6%
Substance abuse indicated by any recorded medical diagnosis or justice system charge	61%	31%
Number of drug abuse and alcohol service encounters in the last 2 years	14.3	3.9
HUD-funded Homeless Services and County Public Assistance		
Chronic homeless flag in any HUD-funded homeless service provider record	27%	11%
Public assistance benefits received this year	46%	40%
Two or more months of food stamp payments received in the past 2 years	47%	44%

Adjusted odds ratios presented in Table 2 reflect the differences we observe from descriptive comparisons. Odds ratios for binary variables (for example, arrested or not) are generally higher than the odds ratios for continuous variables (for example, days in jail) and are interpreted differently. For example, the odds ratios show that persons who have been arrested in the past two years are 1.74 times more likely to be in the high-cost group than those who have not been arrested. On the other hand, the odds ratio for each additional arrest is only 1.06, increasing the likelihood (or odds) of being in the high-cost group by 6 percent.

Odds ratios analysis reveals that being arrested in the last two years, higher jail security and substance abuse are among the strongest binary predictors of becoming a high-cost homeless resident, followed by being arrested for inebriation and released within 48 hours, heart disease, two or more emergency medical service encounters, being admitted as a hospital inpatient via the emergency room, two or more mental health outpatient visits, and receiving public assistance benefits. All factors included in the model increase the likelihood of becoming a high-cost homeless person with adjusted ratios in the range of 1.05 and 1.28, with the exception of receiving two or more months of food stamp payments, which has an odds ratio of 0.68, indicating that receiving food stamps benefits makes it less likely to be in the high-cost group. The adjusted odds ratios for continuous variables all have values ranging from 1.002 (number drug abuse and alcohol services encounters) to 1.16 (number of hospital admissions), and all increase the likelihood of becoming a high-cost homeless person.

General performance of the model was evaluated using C-statistic to assess the predictive ability of the model. The model achieved a very strong C-statistic: .813. C-statistic is the probability that predicting the outcome is better than chance. Models are typically considered reasonable when the C-statistic is higher than 0.7 and strong when C-statistic exceeds 0.8 (Hosmer and Lemeshow, 2000). Overall, the model predicts high-cost homeless persons with a very good fit.

Table 2 Logistic Regression Adjusted Odds Ratios and 95% Confidence Limits for Predictor Variables (Validation Sample)

Variable	Odds Ratio	95% Confidence Limits
Demographics		
Age 18-45 vs. less than 18*	1.21	1.06 – 1.38
Age 46-65 vs. less than 18	.98	.85 – 1.13
Age 65+ vs. less than 18***	.88	.69 – 1.14
Female vs. Male***	1.07	1 -1.14
Criminal Justice		
100+ days of probation in the last 2 years*	1.15	1.03 – 1.28
Arrested in last 2 years*	1.74	1.58 – 1.92
Jail booking in last 2 years*	1.14	1.04 – 1.26
Jail security classification of 3 or 4 (i.e., high risk) this year*	1.63	1.41 – 1.89
Arrested for inebriation and released within 48 hours this year*	1.48	1.26 – 1.73
Number of arrests this year**	1.06	1.01 – 1.11
Number of days in jail this year*	1.007	1.005 – 1.009
Health Diagnoses		
Diagnosed with chronic medical condition*	1.21	1.1 – 1.33
Diagnosed with adjustment reaction in last 2 years*	1.26	1.06 – 1.49
Diagnosed with heart disease in last 2 years*	1.41	1.15 – 1.72
Number of medical encounters with diagnosis of organ failure in last 2 years*	1.08	1.06 – 1.11
Diagnosed with schizophrenia in last 2 years**	1.23	1.03 – 1.46
Number of medical encounters with diagnosis of neoplasm in last two years*	1.05	1.03 – 1.07
Diagnosed with "other ill-defined and unknown causes of morbidity and mortality" in last 2 years **	1.28	1.05 – 1.58
Diagnosed with high-cost ICD-9 in last 2 years**	1.12	1.009 – 1.24
Health & Emergency Services		
Emergency Medical Service (EMS) encounter this year*	1.27	1.14 – 1.41
Emergency Medical Service (EMS) encounter last year*	1.26	1.14 – 1.4
Two or more EMS encounters in last 2 years*	1.34	1.12 – 1.6
Admitted as hospital inpatient via emergency unit admission in last 2 years*	1.35	1.19 – 1.54
Outpatient Psychiatric Emergency Services or ambulatory surgery this year*	1.21	1.11 – 1.33
Number of hospital inpatient admissions this year*	1.16	1..09 – 1.25
Number of hospital inpatient days in last two years*	1.011	1.006 – 1.016
Non-inpatient (ER or clinic) health system encounter this year*	1.2	1.1 – 1.32
Number of non-inpatient (ER or clinic visits) encounters this year*	1.024	1.015 – 1.033
11+ non-inpatient (ER or clinic) health system encounters this year*	1.27	1.07 – 1.51
Behavioral Health		
Number of Mental Health outpatient days in the last 2 years*	1.013	1.01 – 1.015
Two or more Mental Health outpatient visits in the last 2 years*	1.4	1.23 – 1.59
Number of Mental Health inpatient admissions this year*	1.002	1.002 – 1.003
Two or more Mental Health inpatient admission in the last 2 years*	1.28	1.08 – 1.51

Substance abuse indicated by any recorded medical diagnosis or justice system charge *	1.63	1.51 – 1.76
Number of drug abuse and alcohol service encounters in the last 2 years*	1.002	1.002 – 1.002
HUD-funded Homeless Services and County Public Assistance		
Chronic homeless flag in any HUD-funded homeless service provider record*	1.28	1.17 – 1.39
Public assistance benefits received in the current year*	1.36	1.18 – 1.57
Two or more months of food stamp payments received in the past 2 years*	.68	.59 - .79

*p < .01, **p < .05, ***p < .10

Table 3 shows the predictive performance of the model for different scenarios-top one percent, five percent, 10 percent, as well as top 1,000 homeless persons with the highest risk of becoming a high-cost service user. The predictive performance measures were defined earlier in the methods section.

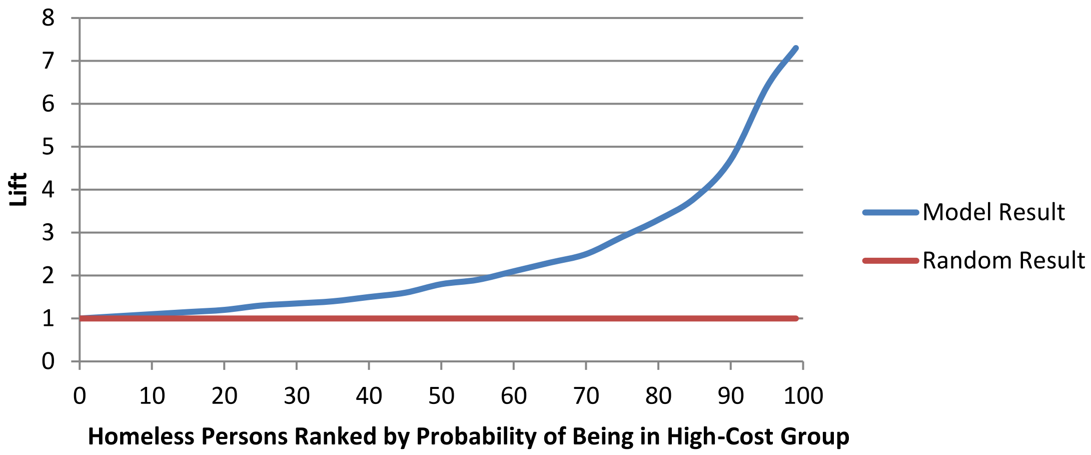
Table 3: Predictive Performance of the Model

Measure	Top 1%	Top 5%	Top 10%	Top 1,000	Formula
Sensitivity	9.3%	32.6%	47.7%	14.9%	True Positive /(True Positive + False Negative)
Specificity	99.7%	97.3%	93.2%	99.4%	True Negative /(False Positive + True Negative)
PPV	72.9%	51%	37.4%	66.8%	True Positive /(True Positive + False Positive)
Accuracy	92.6%	92.3%	89.5%	92.7%	(True Positive + True Negative) /Number

If the top five percent persons (2,864 persons) at risk of becoming high-cost homeless service users are followed, the achieved sensitivity and specificity are 32.6 percent and 97.3 percent, respectively. These values suggest very reasonable predictive power, indicating that the model picks up 33 percent of all high-cost service users and correctly identifies 97 percent of those who are not high users. The PPV value of 51 percent and accuracy value of 92.3 percent for the top five percent are also very high. If we follow a subset within the top five percent, the 1,000 cases with the highest probability scores for being in the high-cost group (1.75 percent of all cases), we see even more accurate prediction outcomes. The model achieves a PPV result of 67 percent, meaning that out of 1,000 persons that model identified as high-cost persons, two-thirds are true positives and the remaining one-third are false positives. PPV is an important measure for assessing the cost-effectiveness of the model.

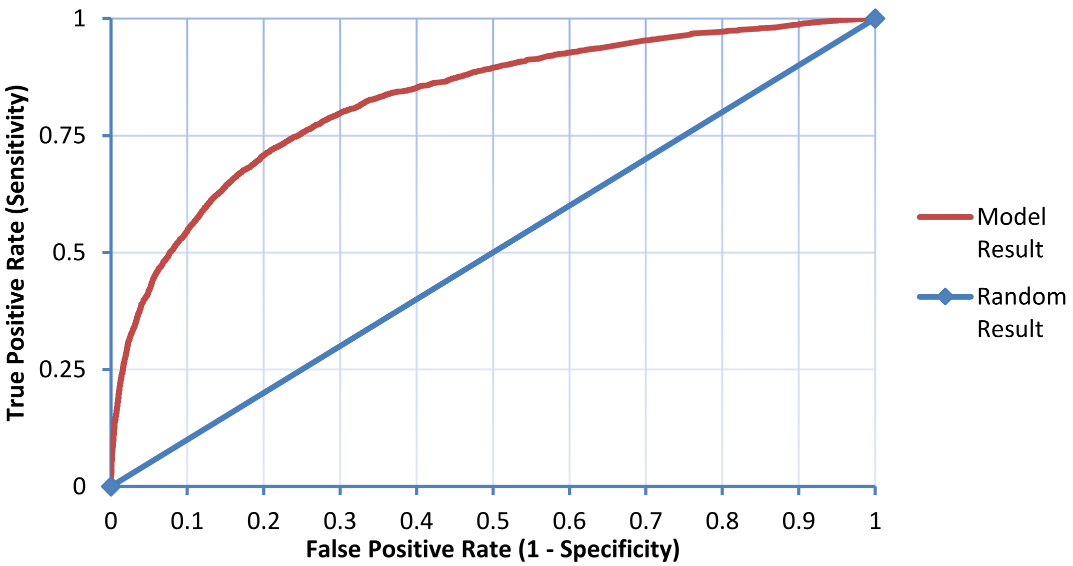
Another measure of the effectiveness of a predictive model is the “lift”, which is calculated as the ratio between the results obtained with and without the predictive model for all thresholds. Figure 1 illustrates the lift of the model, which is quite high for cases with a high probability of being in the high-cost group. For example, for the top five percent, the model generates a lift of 6.5. This means that model generates 6.5 times more correctly identified high-cost homeless persons (true positives) than random selection, which is presented as the baseline-a lift of one or no lift. At slightly lower thresholds, such as the top 10 percent, lift drops to 4.7 because in order to capture more true positives, the model concurrently includes more false positives. Conversely, the number of false positives decreases as the probability of being in the high-cost group increases.

Figure 1 Lift Chart



The most common way of assessing the predictive power of a model in the data mining literature is the area under the ROC curve. ROC shows the trade-off between true positives (sensitivity) and false positives (1-specificity) at all possible thresholds. The ROC curve for the model is shown in Figure 2. The model generated a fairly high AUC of 0.83, indicating an 83 percent probability that a randomly selected homeless person with high future costs will receive a higher

Figure 2 ROC Curve



model score than a randomly selected homeless person without high future service costs. In the predictive analytics literature, models with AUC exceeding .8 are accepted as models with good predictive power while AUC values below 0.7 indicate poor model performance.

Since the model provides a probability score ranging from 0 to 1, we have to select a cut-off score or a threshold to identify who will be offered permanent supportive housing—those homeless persons with scores higher than the selected threshold. Choice of a cut-off level introduces the trade-off between the correct identification of high-cost service users and false alarm rates. The ROC curve illustrates this trade-off between true positives —finding as many homeless persons as possible who would be high-cost service users next year and false positives—decreasing potential cost savings by including homeless persons who would not be high-cost service users next year.

Business Scenario and Cost Savings

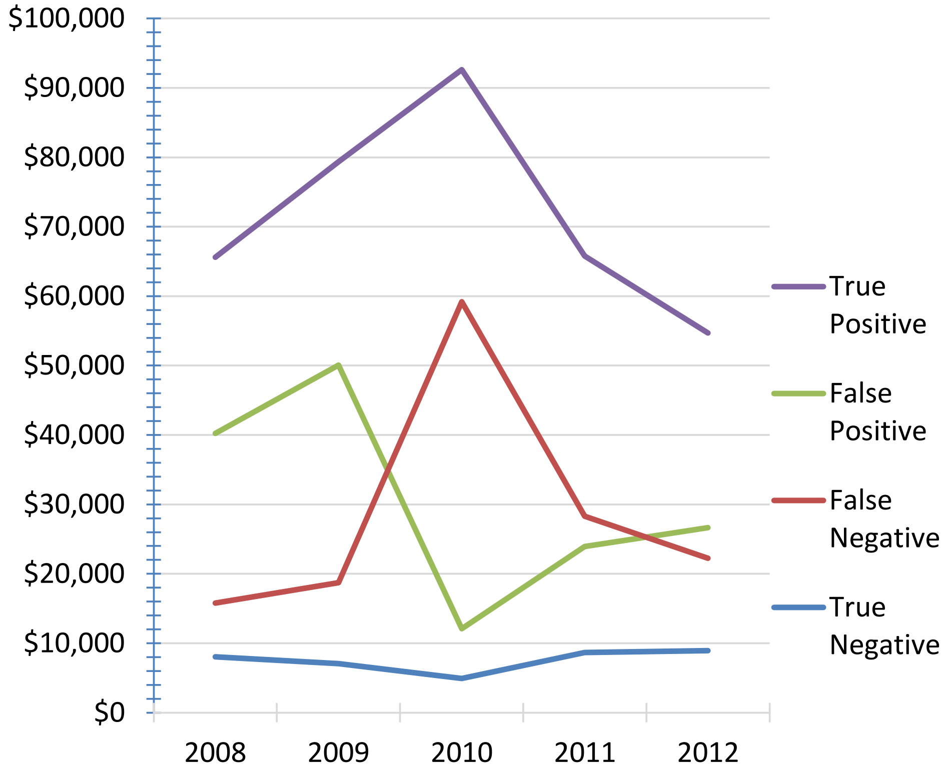
While the performance of the triage tool presented in this paper is very high, it is still necessary to translate this performance into a pragmatic business scenario showing how the tool contributes to the efficiency of PSH programs by prioritizing the population to be housed. The trade-off to be weighed in using the triage tool is between, on the one hand, using lower selection thresholds in order to find as many high-cost homeless individuals as possible but accepting a substantial number of lower-cost individuals as part of the mix, and, on the other hand, using higher selection thresholds to identify a smaller population in which a higher proportion of individuals will be high cost service users. This trade-off is critical to the efficiency of a PSH program as elaborated earlier. The model is highly accurate in distinguishing high-cost from low-cost users, however it is still necessary to calibrate the cut-off level based on goals for saving costs by offering PSH to the targeted population. The following analysis explores the cost efficiency of providing PSH to targeted high-cost homeless persons under different cut-off levels.

Using five years of actual cost data, from 2008 through 2012, it was possible using the first two years of data to produce probability scores for the likelihood of each individual being in the highest-cost group in 2010, and then track the accuracy and financial outcomes of these predictions over the next two years. The efficiency of a PSH program can be evaluated by estimating the cost-offsets. Any placement decision has cost implications. If the homeless person predicted to be a high-cost user was correctly identified (*true positive*), then substantial net cost savings would be observed after accounting for housing and service costs because of the roughly two-thirds reduction in post-housing utilization of public services. However, if the homeless person predicted to be a high-cost user was a *false positive*, then the expected cost savings would not be realized. Instead, the housing and service costs would lead to negative savings. The balance between the positive and negative savings generated by these two groups would determine the efficiency of a PSH program.

One of the challenges the model must contend with is abrupt changes in costs in the scoring year, the year following the two years for which health conditions and service utilization are known. Some conditions are one-time events, resulting in costs that spike and then decline. Hence, the assessment of cost-offsets should be done in the post-placement period, when the actual service utilization of *true positives* and *false positives* becomes evident. Some homeless persons who were *true positives* at the time of scoring year became low-cost users in subsequent years due to regression to the mean. On the other hand, some *false positives* that were predicted to be high-cost users but were low-cost users in the scoring year turned out to have higher costs in subsequent years.

These dynamics are shown in Figure 3. Looking at two years of post-scoring year cost data (adjusted to 2014 dollars), the model successfully differentiates the highest cost cases from other cases even though average costs decline because of regression to the mean. The high specificity of the model is verified by the low-cost levels of *true negatives*. Another critical observation is that public costs for individuals experiencing homelessness vary significantly from one year to the next with important implications for the efficiency measure. *False positives* represent homeless persons with high service utilization prior to the scoring year of 2010, which led to high probability scores. However, in 2010 their service costs were low making them *false positives*. On the other hand, their post-prediction trend is positive, more than doubling between 2010 and 2012.

Figure 3 **Average Annual Costs for Triage Tool Prediction Groups**



It should be also noted that *false negatives,* the group with low service utilization prior to 2010 and high-costs in 2010, the scoring year, typically had one-time cost spikes. Their long-term trend is negative and subsequent to the scoring year their cost levels declined substantially. Hence, omitting them as high-cost users contributes to the efficiency of the program significantly as presented below. Figure 3 suggests that cost savings should be assessed not at the year of scoring but rather in the post-scoring years in order to capture the long-term service utilization of scored individuals.

The triage tool works to assign high scores to high-cost users, but at different probability cut-off levels there will be different proportions of true *positives* with expected savings and *false positives* with no expected cost savings. Our estimation of net savings at different cut-off levels are based on the estimated cost savings for *true positives* after taking into account the housing and service costs for *false positives.* The results are sensitive to the probability score threshold, cost of housing and the rate of anticipated reduction in service utilization and costs following placement in housing. As the probability score threshold increases, the ratio of *true positives* to *false positives* also increases, resulting in increased savings.

This analysis looks at financial outcomes based on two probability score thresholds, 0.37 and 0.53, for the predicted probability of having high costs in 2010, based on 2008 and 2009 information. The 0.37 cut-off level identifies approximately five percent of the test population with complete record linkage data as high-cost users. The 0.53 cut-off level identifies the top 1,000 high-probability service users in our test population. A different probability cut-off can be selected based on the requirements of specific initiatives to address homelessness. If the goal is to house a larger number of high-cost homeless persons, lower cut-off levels may be selected, resulting in lower savings per person. On the other hand, if the supply of housing is limited and a smaller number of high-cost homeless persons can be housed, than a higher cut-off level may be selected, resulting in higher savings per person.

It is assumed that the annual cost of PSH is $17,000 per person per year, based on rent subsidy and supportive service costs in Los Angeles. This high-side cost estimate is based on $11,000 annually for rental subsidy, including first-year costs for temporary housing and benefits advocacy, and $6,000 for supportive services. Actual costs may be lower based on the level of subsidies built into different affordable housing projects and the level of long-term supportive services needed by tenants after they are stabilized in housing. Finally, the post-housing reduction in service costs is assumed to be 68 percent for homeless persons in the 10th decile based on a recent study from Los Angeles (Economic Roundtable, 2009). Most other studies estimate service cost reductions for homeless persons in permanent supportive housing for the whole population, rather than the top decile (Culhane, 2008; Culhane and Byrne, 2010). It is also assumed that there will not be any cost reduction for individuals below the top decile. This is a conservative assumption since an earlier study found post-housing cost reductions among lower-cost individuals (Economic Roundtable, 2009). Hence, net savings are -$17,000 for low-cost user groups because no cost savings are applied to them.

Table 4 presents actual cost savings for 2011 for the two selected cut-off levels (0.37 and 0.53). Post-housing costs savings are calculated as 32 percent of homeless costs for individuals in the 10th cost decile, and then $17,000 is added for each person in the group to cover the cost of housing and supportive services. Net savings are calculated by subtracting estimated post-housing costs from actual homeless costs for the year. All analysis was conducted in 2014 prices. Since actual costs in 2011 and 2012 were used, *regression to the mean*, that is the tendency of extreme outcomes to be closer to the average when measured a second time, has been incorporated into the estimations.

Cost differences were estimated for four probability-cost groups, which each show different cost dynamics. If a score was above the selected cut-off (0.37 or 0.53) and 2010 costs were in the top decile, the record is a *true positive*. However, in subsequent years, *true positives* in 2010 may remain high-cost or become low-cost service users. The long-term cost status of individuals was evaluated based on their actual cost rankings in 2011 or 2012. If they were in the top decile in 2011 or 2012, they were identified as long-term high-cost users. Otherwise, they were identified as low-cost users.

If a score was above the selected cut-off (0.37 or 0.53) and 2010 costs were *not* in the top decile, the record is a *false positive. False positives* may also become high or low cost service users in the future. This was tested by observing actual costs in 2011 and 2012, and identifying cases that moved into the *true positive* cost category. Table 4 shows that at the 0.37 cut-off level, out of the 1,123 individuals who were *true positives*, 255 became low-cost users in 2011. This cost shift was more than offset by 347 *false positives* that turned out to be high-cost users in 2011. In sum, out of 1,889 individuals, 1,115 (60 percent) were high cost users in 2011.

If the five percent (0.37 cut-off level) with the highest probability of being high cost service users were housed permanently with supportive services, savings of over $22 million were estimated in 2011. Even though 40 percent of individuals were low-cost users in 2011 and would not be generating any cost savings, the net savings from the remaining 60 percent shows the feasibility of the intervention. The analysis shows a cost reduction of almost $12,000 per housed homeless person for the top five percent of the population identified by the triage tool as having the greatest probability of high future costs.

The results are even more positive when a higher cut-off level is selected since the accuracy of the tool in predicting high-cost users improves as the probability level increases. The 2011 cost analysis for 1,000 persons in the test population with the highest probability scores, scores at or above 0.53, shows that almost two-thirds (653 individuals) were true positives. Evaluating actual costs in 2011, it is observed that 122 of them became low-cost users, while more than four-fifths, 531, remained high-cost users. In addition, 165 false positives turned out to be high-cost users in 2011. In sum, out of 1,000 individuals, 696 (70 percent) were high cost users in 2011. As expected, the feasibility of the intervention is higher at the 0.53 threshold than at the 0.37 threshold, with an estimated cost reduction for this group of over $19,000 per person in 2011.

Table 4: Cost Savings for 2011 at the Cut-off Levels of 0.37 and 0.53

Status	2010 Costs (Pred. Year)	2011 Costs (1 yr. after Pred.)	2011 Cost Savings	2011 Net Savings	2011 Total Savings	Number Of Cases
Cut-Off Level: 0.37						
True Positives- Low Cost Users	$90,989	$10,932	$0	-$17,000	-$4,335,000	255
True Positives- High Cost Users	$93,196	$83,661	$56,889	$39,889	$30,635,068	768
False Positives- Low Cost Users	$11,444	$8,511	$0	-$17,000	-$8,823,000	519
False Positives- High Cost Users	$13,029	$46,551	$31,655	$14,655	$5,085,204	347
Total / Average				**$11,944**	**$22,562,272**	**1,889**
Cut-Off Level: 0.53						
True Positives- Low Cost Users	$111,580	$11,496	$0	-$17,000	-$2,074,000	122
True Positives- High Cost Users	$96,892	$86,947	$59,124	$42,124	$22,367,823	531
False Positives- Low Cost Users	$12,427	$8,829	$0	-$17,000)	-$3,094,000	182
False Positives- High Cost Users	$13,579	$43,560	$29,621	$12,621	$2,082,432	165
Total / Average				**$19,282**	**$19,282,255**	**1,000**

A separate analysis estimated savings in 2012 for both cut-off levels. Since lower cost levels were observed in 2012 due to the regression to the mean, lower cost savings were estimated. At the 0.37 level, cost savings were estimated to be almost $16 million, which corresponds to over $8,000 per housed individual. At the 0.53 level savings per individual were estimated to be $16,000, with cumulative savings for 2011 and 2012 estimated to exceed $35 million. Over the two years of post-prediction data that we have for Santa Clara County, we see a year-to-year decline in actual costs for individuals with a high probability of having high costs. However, this may be the first phase of a longer-term cost cycle in which costs begin to increase again. This scenario is plausible considering that most individuals in this population have serious medical and mental health disorders that are likely to become more acute as they age. Indications of a longer-term cycle in which costs decline and then increase were found in an earlier cost study in Los Angeles (Economic Roundtable, 2009).

As noted earlier, our cost savings analysis assumed that the annual cost of PSH is $17,000 per person per year and that the post-housing reduction in service costs is 68 percent for homeless persons in the 10th decile. Since both of these assumptions are made based on data and recent studies from Los Angeles, a separate sensitivity analysis was carried out to see how total net cost savings estimates change if these cost assumptions change. The analysis showed that at the 0.37 cut-off level, the break-even point is reached when the annual cost of PSH is $29,000 or the post-housing reduction in service costs is 40 percent. These are the highest annual cost of PSH and the lowest percentage of service cost reduction that still yield net cost savings.

Discussion

This is the first attempt in Santa Clara County and one of the first studies to develop and validate a predictive model for identifying homeless persons who are likely to become high-cost users of public service. This model was developed using an integrated database built by linking seven agencies administrative records, which provided information on risk factors such as demographics, clinical variables and service utilization variables for the current and previous years as well as cost of service data.

An earlier study confirmed the chronic homelessness is very costly to Santa Clara County. The 10 percent with the highest costs, the tenth decile, accounted for 61 percent of all public costs for homelessness and the top five percent accounted for 47 percent of all costs (Economic Roundtable, 2015). The past research showed that permanent supportive housing provided to chronically homeless with relatively higher service costs generated large enough cost offsets to cover the costs of housing and services. However, the number of homeless people needing housing far exceeds the available housing supply, and there has not been a fair, objective system for prioritizing who gets to be housed. Often, the scarce supply of permanent supportive housing is rented out to the eligible population based on crude screening processes that rely on self-reported data. Given that permanent supportive housing is proven to have a large impact on reducing chronic homelessness and associated public costs, there is a strong argument for using more accurate screening tools to identify individuals who should have first priority for access to permanently affordable housing.

The Silicon Valley Triage Tool pulls together fragments of information captured in public records about individuals experiencing homelessness to estimate future public costs and identify people for whom for whom the solution of housing costs less than the problem of homelessness. Presented results suggest that the performance of the model is very strong with high sensitivity and specificity values, and the model was validated for an out-of-sample validation cohort.

The model is particularly strong when using high probability cut-off levels, generating small numbers of false positives and high numbers of true positives. For the top 1,000 high-cost users predicted by the model, two-thirds of them are true positives. A key strength of this study is that it assessed the overall effectiveness of predictions made by the tool, looking at costs over the three years following the two years that were the source of data used to make the prediction. This assessment showed that many false positives became high-cost or close to high-cost users in the second year after the prediction. In addition, a majority of the false negatives were actually true negatives over the next two years because their high cost level in the scoring year represented a one-time cost spike. One of the challenges the model must contend with is abrupt changes in costs from one year to the next. Some conditions are one-time events, resulting in costs that spike and then decline. The tool performed very well by giving low scores to homeless persons with one-time cost spikes.

Another key strength of the study is information it provided for identifying distinctive attributes of high-cost individuals. Individuals in this group are the most likely to be diagnosed with a mental disorder, in particular, a disorder that takes the form of a psychosis, and a psychosis that takes the form of schizophrenia. They are also the most likely to be given a maximum or high-medium security jail classification because of the safety risk they are perceived to present. They are the most likely to have been continuously homeless for three years. They are most likely to be diagnosed with a skin disease such as cellulites or an endocrine disease such as diabetes. They are most likely to be tri-morbid – diagnosed with a mental disorder, a chronic medical condition and to abuse drugs or alcohol. Demographically they are most likely to be male and to be in the middle of their lives - 35 to 44 years old. And they are most likely to frequent users of hospital emergency rooms and inpatient beds, emergency psychiatric facilities, mental health inpatient facilities, and to be incarcerated in a jail mental health cell block.

This composite profile can help hospital and jail discharge planners and homeless service providers identify high-cost individuals. However, there is significant diversity in the demographic attributes and types of crisis services needed by individuals in this population. The triage tool weighs the likely cost impact of each individual's characteristics and uses this information to identify subgroups that fall outside this profile. For example, young women with acute mental illnesses and endocrine diseases who are not substance abusers and not involved in the justice system but are likely to have ongoing high costs.

The model was validated further by developing a business analysis to assess its cost effectiveness. Selecting 0.37 as the optimal cut-off level, which identifies five percent of the population as the target group, the model assessed cost savings by comparing total housing and service costs ($17,000 annually) with the estimated 68 percent cost savings for true positives - those correctly identified as high-cost service users. The results confirmed that anticipated cost savings from true positives far exceed the total costs of housing, yielding net savings of $20,000 per person over the next two years after the total population with a probability score of 0.37 or higher enters permanent supportive housing. Using 0.53 as the minimum probability threshold for the target group, there are estimated annual savings of $32,000 per person after paying for housing and supportive services. On the other hand, using 0.20 as the probability threshold, we achieve break-even financial results, with cost savings from reduced service use fully offset by the cost of providing housing and supportive services.

The optimal cut-off is not simply an empirical decision. In the context of permanent supportive housing it depends on the number of people who can be housed in available housing. However, in the context of a long-term strategy to address homelessness, the trade-off between costs and savings in the population needing housing provides evidence that jurisdictions can use to validate initiatives such as affordable housing bond measures to expand the inventory of available housing.

It is often argued that the feasibility of prevention services such as permanent supportive housing would not be attained without a strategy of balancing the costs with some degree of cost offsets. One of the most significant strengths of this study is its strong performance in identifying homeless persons with high probability of having high ongoing public costs that will substantially exceed the cost of permanent supportive housing.

The predictive performance of the Silicon Valley Triage Tool was compared to the performance of two earlier triage tools developed in Los Angeles by running all of the models on records of homeless persons from both Los Angeles and Santa Clara counties. The tools were assessed based on the proportion of high-cost homeless persons correctly identified by each model and the proportion of persons predicted to be high-cost homeless who truly were high-cost persons. The Silicon Valley tool demonstrated comparable or higher accuracy when run on Los Angeles data and much higher accuracy when applied to the Santa Clara data. This comparison verifies that the Silicon Valley tool demonstrates strong predictive performance in multiple metropolitan regions.

Limitations

This analysis and the model developed in this study are also subject to some limitations that need to be acknowledged and most of these limitations are inherent to analysis involving administrative data sets. Our study was limited by the usual shortcomings of research based on linked administrative records, including errors in the underlying data sources, such as missing data and data entry errors. Matching inaccuracies prevented the use of the full homeless population for the analysis. Roughly 55 percent of the population, 57,259 homeless persons, were used to develop the tool. These were individuals with at least one record linked to an agency during our six-year study window from 2007 through 2012. Since administrative databases do not collect data for research purposes, some of the critical risk factors were not available such as the income and employment of homeless persons. Moreover, some service costs were missing for some years and had to be estimated. For some services, when individual-level costs were not available, average costs per unit of service were used.

Another shortcoming related to the use of administrative data is incomplete and sometimes inaccurate information about the timing of homeless episodes. Since complete information about the duration of homelessness was not available, the study population was assumed to be either homeless or at risk of homelessness while predicting high-cost users, assuming that individuals would use more services when they were experiencing homelessness. In addition, the administrative datasets did not show the mobility of homeless individuals in and out of the county, which would impact their utilization of services in county facilities.

The business scenario that estimated cost savings was also subject to some limitations. First, it assumed that PSH costs $17,000 a year, which needs to be verified when the county has a larger body of post-supportive housing cost data. Second, since post-housing costs of homeless persons were not available for this study, cost offsets were based on a saving factor of 68 percent, which was derived from an earlier study conducted in Los Angeles. Actual cost savings may be different after the implementation of the program. On the other hand, service reductions measured here represent a conservative assessment of the impact of the PSH on service use and costs because it was assumed that homeless persons with costs below the 10th decile would not experience any service reductions after being housed, so that PSH costs were not adjusted with any cost-offsets for this group.

Finally, the Silicon Valley Tool is a system-based tool, that is, it requires detailed health care and justice system information about each individual that is available only from those institutional systems. This includes medical diagnoses, accurate details of encounters with health care providers, and details about stints of incarceration. Cooperation of both health care and justice system agencies is necessary to obtain information required for the tool.

Because of the level of effort required to obtain and integrate the necessary data, the most efficient use of the tool is for regular, ongoing system-wide screening of linked records rather than screening clients individually. By predicting how likely each person in the entire identified population of homeless resident is to have high future costs, it is possible to prioritize individuals for access to the scarce supply of permanent supportive housing. For example, targeted individuals can be flagged in client databases so that housing can be offered to them the next time they seek services.

The Silicon Valley Tool can also be used to screen cases individually. A version of the tool for individual screening in Excel format as well as software code for screening entire client databases can be downloaded at www.economicrt.org.

Because the tool does not correctly identify all high-cost individuals, the screening process for either individuals or groups should include an option to over-ride the triage tool probability score based on the clinical judgement of health care professionals. For example, if a patient has recently been diagnosed with a high-cost, chronic medical condition, this would warrant overriding a negative result from the triage tool and including the patient in the high-cost group that receives access to permanent supportive housing. Allowing overrides permits service providers to adapt to changing populations and conditions and to react to unique circumstances.

The tool also has practical value for identifying patients served by health plans and private hospitals who have high ongoing costs, and whose health outcomes will improve and costs decrease if they are housed. Local government safety net resources can be augmented through collaborative care for frequent users who are also served by private hospitals.

Using the triage tool raises the broader ethical issue of making decisions about who gets into housing and who is left out. We see the tool as an interim means of prioritizing need in the context of social failures to provide an adequate supply of affordable housing or to provide more effective social safety net interventions that will reduce the flow of people into chronic homelessness. In this context, the tool prioritizes individuals based on public costs, which reflect frequency of service-intensive crises, and are closely linked to (but not identical with) level of distress. Use of the triage tool may be the approach that houses the greatest number of people because public agencies achieve the highest level of cost avoidance by housing high-cost individuals, opening the possibility using those savings to subsidize housing for a larger pool of homeless people.

Conclusion and Future Research

Needs within the homeless population vary significantly. While the Silicon Valley Triage Tool is effective for prioritizing access to permanent supportive housing for the small number of high-cost individuals who account for the majority of public costs, other tools are needed to target services for less disabled segments of the population. Permanent supportive housing is expensive and scarce. Less expensive interventions are effective for individuals with less acute needs. Without effective early intervention there is a real risk that individuals will become chronically homeless, and even that their problems will worsen to the extent that they become high-cost homeless.

Additional predictive tools are needed to effectively target segments of the population experiencing homelessness that are appropriate for earlier interventions. This includes preventive care for children, who have experienced homelessness, integrated outpatient health care, readily available and effective mental health services, temporary affordable housing, and employment services. Immediate employment assistance for employable individuals is essential because re-entering the labor market becomes increasingly difficult the longer individuals are disconnected from work, and, for many individuals, employment is a genuine possibility for escaping acute poverty and homelessness.

The Silicon Valley Triage Tool is the first of multiple triage tools that are needed to target services that are cost effective in meeting the needs of different segments of the homeless population. While PSH may not be appropriate or cost-effective for every person who is homeless, it is a crucial resource for high-need individuals whose post-housing cost reductions can offset the costs of the program. The tool demonstrates that predictive risk models can offer a substantial bonus in efficiency in homelessness prevention services, connecting services to people most likely to benefit from them.

Acknowledgements

This research was underwritten by Destination: Home and the County of Santa Clara.

References

Apicello, J. (2010). A paradigm shift in housing and homeless services: Applying the population and high-risk framework to preventing homelessness. *The Open Health Services and Policy Journal*, *3*, 41-52.

Bcom, A.L. and Latimer, E. (2015). Housing first impact on costs and associated cost offsets: A review of the literature. *Canadian Journal of Psychiatry*, *60,* 275-87.

Burt, M. R. (2002). Chronic homelessness: Emergence of a public policy. *Fordham Urban Law Journal*, *30*, 1267-1279.

Burt, M.R., Pearson, C.L., Urban Institute and Walter R. McDonald and Ass. (2005). *Strategies for preventing homelessness.* Washington, DC: HUD.

Byrne, T., Fargo, J. D., Montgomery, A. E., Munley, E., and Culhane, D. P. (2014). The relationship between community investment in permanent supportive housing and chronic homelessness. *Social Service Review*, *88*, 234–263. doi:10.1086/676142

Byrne, T., Treglia, D., Culhane, D. P., Kuhn, J. and Kane, V. (2016). Predictors of homelessness among families and single adults after exit from homelessness prevention and rapid re-housing programs: Evidence from the department of veterans affairs supportive services for veteran families hrogram" *Housing Policy Debate*, *26*, 252-275. doi:10.1080/10511482.2015.1060249.

Caton, L. M., Dominguez, B. D., Schanzer, B., Hasin, D.H., Shrout, P.E., Felix, A., Mc Quiston, H., Opler, L.A. and Hsu, E. (2005). Risk factors for long-term homelessness: Findings from a longitudinal study of first-time homeless single adults. *American Journal of Public Health*, *95*, 1753-1759. doi: 10.2105/AJPH.2005.063321

Caton, C., Wilkins, C., and Anderson, J. (2007). *People who experience long-term homelessness: Characteristics and interventions*. Retrieved from http://aspe.hhs.gov/hsp/homelessness/symposium07/caton.

Chechulin, Y., Nazerian, A., Rais, S. and Malikov, K. (2014). Predicting patients with high risk of becoming high-cost healthcare users in Ontario (Canada). *Healthcare Policy*,*9*, 68-79. doi:10.12927/hcpol.2014.23710

Culhane, D. P., Metraux, S. and Hadley, T. (2002). Public service reductions associated with placement of homeless persons with severe mental illness in supportive housing. *Housing Policy Debate***,** 13, 107–163. doi:10.1080/10511482.2002.9521437

Culhane, D. P. (2008). The cost of homelessness: A perspective from the United States. *European Journal of Homelessness***.** *2*, 97-114. Retrieved from http://repository.upenn.edu/spp_papers/148

Culhane, E.P., Metraux, S. and Byrne, T. (2011). A prevention-centered approach to homelessness assistance: a paradigm shift? *Housing Policy Debate*, *21*, 295-315. doi:10.1080/10511482.2010.536246

Culhane, D. P. and Byrne, T. (2010). Ending chronic homelessness: Cost-effective opportunities for interagency collaboration." *Penn School of Social Policy and Practice Working Paper*, Retrieved from http://repository.upenn.edu/spp_papers/143

Economic Roundtable (2009). *Where we sleep: The costs of housing and homelessness in Los Angeles.* doi: 10.13140/RG.2.1.2624.0887

Economic Roundtable (2011). *Crisis indicator: Triage tool for identifying homeless adults in crisis*. doi: 10.13140/RG.2.1.4788.8246

Economic Roundtable (2012). *Hospital to home: Triage tool II for identifying homeless hospital patient in crisis. Los Angeles, CA. Author.*

Economic Roundtable (2015a). *All alone: Antecedents of chronic homelessness*. doi: 10.13140/RG.2.1.4067.9281

Economic Roundtable (2015b). *Home not found: The cost of homelessness in Silicon Valley.* Retrieved from doi: 10.13140/RG.2.1.4780.6327

Federal register (2015) Homeless emergency assistance and rapid transition to housing: Defining ''Chronically Homeless''. *80*, 75791-75806. Retrieved from https://www.hudexchange.info/resource/4847/hearth-defining-chronically-homeless-final-rule/

Fleishman, H.A. and Cohen, J.W. (2010). Using information on clinical conditions to predict high-cost patients. *Health Services Research, 45*, 532–552. doi: 10.1111/j.1475-6773.2009.01080.x

Folsom et. al. (2005). Prevalence and risk factors for homelessness and utilization of mental health services among 10,340 patients with serious mental illness in a large public mental health system. *American Journal of Psychiatry, 162*, 370-376. doi:10.1176/appi.ajp.162.2.370

Gilmer, T., Willard, P., Manning, G. and Ettner, S. L. (2009). A cost analysis of San Diego County's REACH program for homeless persons. *Psychiatric Services,60***,** 445–450.

Gonen, M. (2007). *Analyzing receiver operating characteristics with SAS*. Cary, NC: SAS Press Series.

Greenwood, R. M., Stefancic, A and Tsemberis, S. (2013). Pathways housing first for homeless persons with psychiatric disabilities: Program innovation, research, and advocacy. *Journal of Social Issues*, *69*, pp: 645-663. doi:10.1111/josi.12034

Henwood, B.F., Dichter, H., Tynan, R., Simiriglia, C., Boermer, K. and Fussaro, A. (2015). Service use before and after the provision of scatter-site Housing First for chronically homeless individuals with severe alcohol use disorders. *International Journal of Drug Policy, 26,* 883-886. doi:10.1016/j.drugpo.2015.05.022

Hosmer, D.W. and Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: John Wiley and Sons.

HUD (2009). *Notice of allocations, application procedures, and requirements for homelessness prevention and rapid re-housing program grantees under the American Recovery and Reinvestment Act of 2009*. Washington, DC. Retrieved from https://portal.hud.gov/hudportal/documents/huddoc?id=hrp-notice.pdf

HUD (2014). *The Annual Homeless Assessment Report (AHAR) to Congress: Volume II*. Washington, DC. Retrieved from https://www.hudexchange.info/resource/4828/2014-ahar-part-2-estimates-of-homelessness/

Kuhn, R, and Culhane, D. P. (1998). Applying cluster analysis to test a typology of homelessness by pattern of shelter utilization: results from the analysis of administrative data. *American Journal of Community Psychology***,** *26,* 207–232. Retrieved from http://repository.upenn.edu/spp_papers/96

Kuno, E, R., Rothbard, A. B., Averyt, A.B. and Culhane, D. P. (2000). Homelessness among persons with serious mental illness in an enhanced community-based mental health system. *Psychiatric Services*, *51*, 1012–1016. doi: 10.1176/appi.ps.51.8.1012

Kushel, M.N., Perry, S., Bangsberg, D., Clark, R. and Moss, A.R. (2002). Emergency department use among the homeless and marginally housed: results from a community-based study. *American Journal of Public Health*, *92***,**778-784. doi: 10.1186/s13722-015-0038-1

Kushel, M.N., Hahn, J.A., Evans, J.L, ., Bangsberg, D. and Moss, A.R. (2005). Revolving doors: imprisonment among the homeless and marginally housed population. *American Journal of Public Health*, *95,* 1747-1752. doi: 10.2105/AJPH.2005.065094

Larimer, M.E. et.al. (2009). Health care and public service use and costs before and after provision of housing for chronically homeless persons with severe alcohol problems. *Journal of American Medical Association*, *301,* 1349-1357. doi:10.1001/jama.2009.414

Larson, E.B. (2013). Building trust in the power of "big data" research to serve the public good. *Journal of American Medical Association*, *309,* 2443-2444. doi:10.1001/jama.2013.5914

Martinez,T. and Burt, M. R. (2006). Impact of permanent supportive housing on the use of acute care health services by homeless adults." *Psychiatric Services*, *57,* 1-8. doi: 10.1176/ps.2006.57.7.992

McNiel, D. E., Binder, R. L., and Robinson, J. C., (2005). Incarceration associated with homelessness, mental disorder, and co-occurring substance abuse. *Psychiatric Services*, *56*, 840–846. doi: 10.1176/appi.ps.56.7.840

McLaughlin T. C. (2011). Using common themes: Cost-effectiveness of permanent supported housing for people with mental illness." *Research on Social Work Practice*, *21*, 404-411. doi: 10.1177/1049731510387307

Metraux, S., Culhane, D., Raphael, S., White, M., Pearson, C., Hirsh, E., et al. (2001). Assessing homeless population size through the use of emergency and transitional shelter services in 1998: Results from the analysis of administrative data from nine U.S. jurisdictions. *Public Health Reports*, *116*, 344–352. Retrieved from http://repository.upenn.edu/spp_papers/85

Metraux S. and Culhane D. P. (2004). Homeless shelter use and reincarceration following prison release: assessing the risk. *Criminal Public Policy, 3,* 201–222. Retrieved from http://repository.upenn.edu/spp_papers/116

Montgomery, A.E., Fargo, J, D., Byrne, T. H., Kane, V. and Culhane D. P. (2013). Universal screening for homelessness and risk for homelessness in the Veterans Health Administration. *American Journal of Public Health, 103,* S210-S211. doi: 10.2105/AJPH.2013.301398

Moturu, A. T., Johnson, W. G. and Liu, H. (2010). Predicting future high-cost patients: A real-world risk modeling application. I*nternational Journal of Biomedical Engineering and Technology*. *3,* 114–132. doi: 10.1504/IJBET.2010.029654

Poulin, S.R., Maguire, M., Metraux, S. and Culhane, D.P.(2010). "Service use and costs for persons experiencing chronic homelessness in Philadelphia: A population-based study." *Psychiatric Services*, *61*, 1093-1098. doi: 10.1176/ps.2010.61.11.1093

Rog, D. J., Marshall, T., Dougherty, R. H., Preethy, G., Daniles. A.S., Ghose, S. S. and Delphin-Rittmon, M. E (2014). "Permanent supportive housing: Assessing the evidence" *Psychiatric Services*, *65***,** 287-294. doi:10.1176/appi.ps.201300261

Rosenheck, R. (2000) "Cost-effectiveness of services for *mentally ill homeless people: The application of research to policy and practice." American Journal of Psychiatry*, *157*, 1563-1570. doi: 10.1176/appi.ajp.157.10.1563

Sadowski, L. S., Romina A., Kee, T. J., Weele, V. and Buchanan, D. (2009). Effect of a housing and case management program on emergency department visits and hospitalizations among chronically ill homeless adults: A randomized trial. *Journal of the American Medical Association*, *301*, 1771–1778. doi:10.1001/jama.2009.561

Sarma, K. S. (2013). *Predictive modeling with SAS Enterprise Miner: Practical solutions for business applications*, NC: SAS Institute.

SAS (2103). *Getting started with SAS Enterprise Miner 13.1*, Cary, NC: SAS Institute.

Shinn, M., Baumohl, J. and Hopper, K. (2001). The prevention of homelessness revisited. *Analyses of Social Issues and Public Policy*, *1*, 95–127. doi:10.1111/1530-2415.00006

Shinn, M., Greer, A. L., Bainbridge, J., Kwon, J. and Zuiderveen (2013). Efficient targeting of homelessness prevention services for families *American Journal of Public Health, 103*, S324-S330. doi: 10.2105/AJPH.2013.301468

Tamang et. al. (2015). Improving the foundation of population-based spending arrangements by predicting "cost blooms". CA: Stanford University. Retrieved from http://statweb.stanford.edu/~ljanson/papers/Cost_Blooms-Tamang_ea-2015.pdf

Toros, H.and Stevens, M. (2012). Project 50: The cost effectiveness of the permanent supportive housing model in the skid row section of Los Angeles County. Los Angeles: County of Los Angeles, CEO.

Tsemberis, S., and Eisenberg, R. (2000). Pathways to housing: Supported housing for street-dwelling homeless individuals with psychiatric disabilities. *Psychiatric Services**, 51,*** 487–493. doi: 10.1176/appi.ps.51.4.487

USICH (2010) "Opening Doors: The Federal Strategic Plan to Prevent and End Homelessness." US Interagency Council on Homelessness, Washington, DC. Author.

USICH (2015) "Opening Doors: The Federal Strategic Plan to Prevent and End Homelessness." Washington, DC. Retrieved from https://www.usich.gov/opening-doors

USICH (2016) "The President's 2016 Budget: Fact Sheet on Homelessness Assistance." US Interagency Council on Homelessness, Washington, DC. Retrieved from https://www.usich.gov/resources/uploads/asset_library/2016_Budget_Fact_Sheet_on_Homelessness_Assistance.pdf

Zugazaga, C. (2004). Stressful life event experiences of homeless adults: A comparison of single men, single women, and women with children. *Journal of Community Psychology*, *32,* 643–654. doi: 10.1002/jcop.20025