Silicon Valley Triage Tool

Identifying and Housing High-Cost Homeless Residents

February 2016

ECONOMIC
ROUNDTABLE

Knowledge for the Greater Good

SILICON VALLEY TRIAGE TOOL

Identifying and Housing High-Cost Homeless Residents

February 2016

Economic Roundtable

Halil Toros

Daniel Flaming

Table of Contents

Executive Summary

Why the Silicon Valley Triage Tool is Important

The number of homeless people needing housing far exceeds the available housing supply, and there is not a fair, objective system for prioritizing who gets to be housed. The triage tool addresses this problem by identifying individuals for whom the solution of housing costs less than the problem of homelessness.

The tool can be used to screen people individually or to screen a large database of people experiencing homelessness. When a large group of people are screened, they can be ranked by their probability of being high-cost users of public services. This is a very effective way of matching people with the highest costs with the available supply of housing. For example, if 1,000 housing units are available, the 1,000 homeless individuals with the highest probability of having ongoing high costs can be placed in those units.

What the Triage Tool Does

People living without addresses become invisible. The fragments of their lives captured in public records remain unconnected. The Silicon Valley Triage Tool pulls these fragments together to provide a picture of future public costs in order to identify people for whom the solution of housing costs less than the problem of homelessness.

The tool uses 38 pieces of information to calculate the probability that a homeless individuals will have ongoing high costs. The information includes age, gender, time spent in jail, medical diagnoses, and use of hospital facilities. The tool produces a probability from 0.01 to 0.99 that an individual will have continuing high public costs.

This predictive analytic tool is the most accurate screening software that has ever been developed to predict future costs of individuals whose lives and struggles change dramatically from one year to the next. It enables public agencies and housing providers to give first priority for housing and supportive services to individuals who are in greatest distress and who will be the most frequent users of public services.

Severely disabled individuals experiencing homelessness usually receive care in hospital emergency rooms, emergency psychiatric facilities, hospital inpatient beds, or jail mental health facilities, but after they are released they often do not receive ongoing primary care and behavioral health services that will prevent recurring crises.

Repeated use of crisis services is extremely costly but can be prevented. The Silicon Valley Triage Tool identifies the most costly users of public services so that they can receive appropriate and cost-effective services together with permanent supportive housing.

The tool has the same 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. County safety net resources can be augmented through collaborative care for frequent users who are also served by private hospitals

Accuracy of the Tool

The biggest job the triage tool does is to correctly exclude most people with low costs. After that, it is distinguishing subtle differences between people with significant problems and significant levels of service use to predict who will have high future costs. 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. A key strength of this tool is that it is based on service records over a five-year period, so the source data captures many of these one-time cost spikes.

As with all predictive modeling algorithms, there are trade-offs between maximizing the number of high-cost homeless persons correctly identified and minimizing the number of low-cost homeless persons incorrectly identified. The model is particularly strong when using high probability cut-off levels. For the top 1,000 high-cost users predicted by the model, two-thirds of them are correctly identified as having high future costs.

We completed our validation by developing a business analysis to assess the cost effectiveness of the model. We identified as the optimal cut-off level a probability of 0.37 or higher of being in the high-cost group. At this probability level, 5 percent of the homeless population is identify as the target group. We assessed potential cost savings by comparing total housing and service costs ($17,000 annually) with the estimated 68 percent post-housing 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.

We extended our cost analysis to different probability cut-off levels because the threshold can be raised or lowered depending on policy objectives and availability of housing. We showed that at higher probability thresholds, per capita savings increase because a higher proportion of high-cost users are targeted. 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.

Predictive Performance in Multiple Metropolitan Regions

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.

Distinctive Attributes of High-Cost Individuals

High-cost individuals are differentiated by having acute problems and using expensive services more frequently than in any other group. 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 the most likely to be tri-morbid – diagnosed with a mental disorder, a chronic medical condition and to

abuse drugs or alcohol. Demographically they are the most likely to be male and to be in the middle of their lives – 35 to 44 years old. And they are the 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.

Cost Effective Intervention

People experiencing homelessness usually receive care in hospital emergency rooms or hospital inpatient beds, but after they are released from the hospital they often do not receive ongoing comprehensive primary care and behavioral health services that would prevent recurring crises. Vulnerable homeless individuals end up in a cycle that includes emergency rooms, detox or sobering centers, ambulances, jails, shelters, and the streets which is distressing for these individuals and costly for the public.

Repeated use of crisis services is extremely costly but can be prevented. This tool identifies the most costly users of public services so that they can be housed permanently and engaged in more appropriate and cost-effective services. These interventions typically result in housing stability, significant reductions in the utilization and costs of emergency health services, and better treatment for complex health issues.

While evidence-based research supports the benefits of permanent supportive housing, the supply is far smaller than the population of homeless persons who need this combination of permanently affordable housing and supportive services. Homeless intervention strategies need to prioritize access to housing based on objective criteria. Often, the scarce supply of permanent supportive housing is rented out to the eligible population of homeless persons on a first-come, first-served basis without any system of priorities. This approach tends to favor less vulnerable homeless persons who do not have difficulties with the application and documentation process for this housing. Needs within the homeless population vary significantly and a small number of individuals experiencing homelessness account for the majority of public service costs. 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 triage tools to identify individuals who should have first priority for access to permanent supportive housing.

Efficient Use of the Tool

This 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.

Overview

Purpose of Triage Tool

This report presents a new, more accurate 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 tool enables these gatekeeper institutions to identify more accurately individuals experiencing homelessness whose acute needs create the greatest public costs, and to make credible requests to housing providers that these individuals be given first priority for the scarce supply of permanently affordable housing with services (permanent supportive housing).

Related research studies consistently find that a large share of public service costs such as hospital inpatient and emergency health services or medical and mental health services in custody are borne by a small group of persons with extraordinarily high costs. A second consistent finding is that a disproportionate share of these high-cost patients are experiencing homelessness.

U.S. data from the Arizona Health Care Cost Containment System showed that 10 percent of patients accounted for two-thirds of healthcare costs (Moturu et al. 2010). Another U.S. study found that 5 percent of the population accounted for 49 percent of total healthcare spending (Center for Healthcare Research and Transformation 2010). In Canada, 5 percent of healthcare users consumed 61 percent of hospital and home care spending (Rais et al. 2013). 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 5 percent accounted for 47 percent of all costs (Economic Roundtable, 2015).

It is very difficult for people with complex health issues experiencing homelessness to access the health care and treatment services they need in an effective and efficient way that results in long-term stability. People experiencing homelessness usually receive care in hospital emergency rooms or hospital inpatient beds, but after they are released from the hospital they often do not receive ongoing comprehensive primary care and behavioral health services that would prevent recurring crises. Vulnerable homeless individuals end up in a cycle that includes emergency rooms, detox or sobering centers, ambulances, jails, shelters, and the streets which is distressing for these individuals and costly for the public.

Repeated use of crisis services is extremely costly but can be prevented. The triage tool presented in this report identifies the most costly users of public services so that they can be engaged in more appropriate and cost-effective services in coordination with permanent supportive housing. These interventions typically result in housing stability, significant reductions in the utilization and costs of emergency health services, and better treatment for complex health issues (Economic Roundtable, 2013). A number of earlier studies have demonstrated that declining chronic homelessness along with cost reductions and service efficiencies can be achieved through integrated deployment of housing, supportive services, and ongoing case management for homeless adults suffering with varied combinations of physical health problems, substance abuse issues, and mental illness (Byrne, et al. 2014; Culhane, Metreux and Hadley, 2002; Economic Roundtable, 2009; Gilmer, Wilard and Ettner, 2009; Rog, et al. 2014; Sadowski, et al. 2009; Toros and Stevens, 2012).

While evidence based research supports the benefits of permanent supportive housing, the supply is far smaller than the population of homeless persons who need this combination of permanently affordable housing and supportive services. Homeless intervention strategies should not only be effective but also efficient—targeting individuals based on objective criteria (Burt et al. 2005). Often, the scarce supply of permanent supportive

Repeated use of crisis services is extremely costly but can be prevented.

----

housing is rented out to the eligible population of homeless persons on a first-come, first-served basis without any system of priorities. This approach tends to favor less vulnerable homeless persons who do not have difficulties with the application and documentation process for this housing. Needs within the homeless population vary significantly and a small number of individuals experiencing homelessness account for the majority of public service costs. 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 triage tools to identify individuals who should have first priority for access to permanent supportive housing.

The Silicon Valley Triage Tool improves on earlier tools developed by Economic Round 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). These earlier papers reported on the exceptionally high public costs for homeless individuals in the 10th decile. This discovery has led to interest in identifying these individuals and giving them top priority for permanent supportive housing. There are *four* reasons for this interest. *First*, individuals in the 10th cost decile have very high public costs, accounting for 61 percent of all public costs for homeless adults in Santa Clara County. *Second*, there are very large cost savings when homeless individuals obtain permanent supportive housing along with the safety and stability it provides. Public costs for individuals in the 10th decile decrease by over 60 percent when they live in permanent supportive housing (Economic Roundtable, 2009; Toros and Stevens, 2012). *Third*, these individuals often need special efforts on their behalf to gain access to permanent supportive housing. This is because majority of this group suffers from mental disabilities that are a barrier to completing multiple detailed applications for benefits and housing, as well as documenting their personal identity, income and disability status. Furthermore, fair housing laws are often interpreted to require providers to make affordable housing available on a first-come, first-served basis, creating a barrier to prioritizing high-need individuals. *Fourth*, high public costs are the result of ongoing crises in individuals' lives that are resolved in expensive institutional settings – jails and hospitals. Increasing the level of stability and reducing the frequency and severity of crises through permanent supportive housing greatly improves the quality of these individuals' lives.

It was possible to develop this tool was made possible by 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. Because this data reliably represents a complete population of homeless individuals, it is possible to rank the public costs for a given homeless individual against the overall population of adults experiencing homelessness.

Summary of Triage Tool Results

Santa Clara County data from 2008 and 2009 was used to predict whether individuals would be in the high cost group in 2010. This makes it possible to see three years of post-prediction outcomes based on actual costs in 2010 through 2012. Records used for this test (but not the statistical analyses in the rest of the report) were filtered to exclude cases that had a status change during the prediction window that was outside the scope of the model. These were individuals who died, were sentenced to prison, transitioned from childhood to adulthood, left the foster care system, or left Santa Clara County.

A probability cut-off of 0.37 was used as the primary breakpoint for identifying people likely to be in the high-cost group in 2010. We also tested several higher breakpoints to

A small number of individuals experiencing homelessness account for the majority of public service costs.

Figure 1: Prediction Results from Using 2008-2009 Data to Predict High-Cost Status in 2010
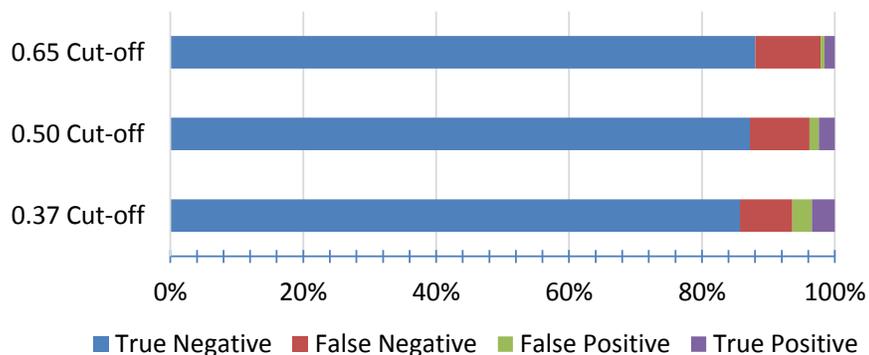
The biggest job the model is doing is correctly excluding most people with low costs. After that, it is distinguishing subtle differences between people with significant problems.

explore the effects of using more stringent probability thresholds for selecting the target population.

Results from these tests are shown in *Figure 1*, which displays the match between predicted and actual cost category in 2010, using different probability cut-off levels for including people in the high-cost group. The high-cost group represents the most costly 10 percent of test sample with complete cost data, and roughly 5 percent of all homeless cases, including records with missing data for which service utilization was estimated using an imputation methodology described in the earlier report, Home Not Found (Economic Roundtable 2015). The average cost in 2010 for the high-cost group was $69,405. The low end of the cost range for this group was $25,362 and the high end was $1,666,872.

The biggest job the model is doing is correctly excluding most people with low costs. After that, it is distinguishing subtle differences between people with significant problems and significant levels of service use to predict who will have high future costs. *Figure 1*

Figure 2: Average Annual Costs for Triage Tool Prediction Groups

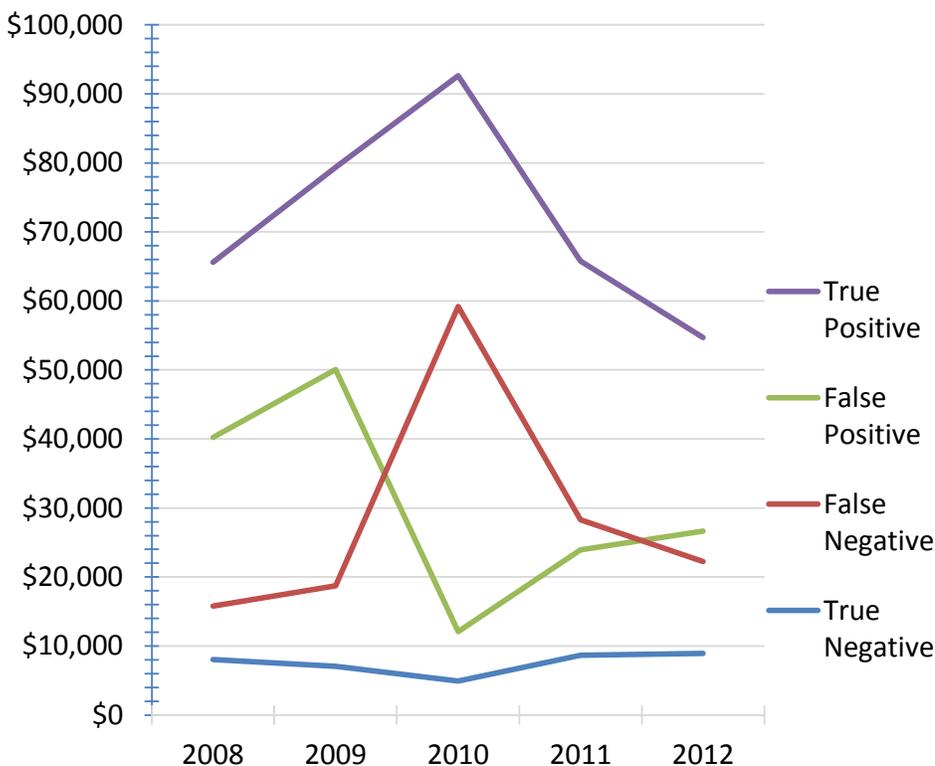shows that there are trade-offs between using different cut-off levels to select people for the high-cost group. Higher cut-offs result in selecting a smaller, higher cost target population by correctly including a higher proportion of people who turn out to actually have high costs. But the unintended consequence is that a higher proportion of people who belong in the high-cost group are incorrectly excluded. These trade-offs are explored in greater depth later in the report. A key issue in choosing the cut-off level is the intended size of the target group, that is, the number of homeless individuals who can be housed.

The results reported in this section are based on using a probability cut-off level of 0.37. Of the cases scored, 6.4 percent were at or above the 0.37 probability cut-off, representing the combined total of 3.5 percent that were true positives and 2.9 percent that were false positives.

The overall results at this cut-off level were that the predictions produced by the model yielded 85.7 percent true negative cases correctly excluded from the high-cost group; 7.9 percent false negatives, records incorrectly excluded from the high-cost group; 2.9 percent false positives, records incorrectly included in the high-cost group; and 3.5 percent true positives, records correctly included in the high-cost group. Looking just at 2010 costs, the model made correct predictions for 89 percent of the cases and incorrect predictions in 11 percent of the cases. The effectiveness of the model becomes clearer when we look at costs for each of these four groups from 2010 through 2012.

One of the challenges the model must contend with is abrupt changes in costs in the year after 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. On the other hand, if individuals remain homeless, problems tend to worsen over time. Looking at three years of post-prediction cost data (adjusted to 2014 dollars) in *Figure 2*, the model successfully differentiates the highest cost cases from other cases. The high-cost group identified by the model has costs in 2010 through 2012 that are more than double the next highest group, the false negatives.

One of the challenges the model must contend with is abrupt changes in costs.

Probability Distribution of Low- and High-Cost Cases

The four classes of records we use to assess the validity of the Silicon Valley Triage Tool are introduced in *Figure 3*. We show the probability scores from 0 to 100 in the horizontal axis. The vertical axis is a partial scale showing the number of low- and high-cost cases with each probability score.

The total number of *true negatives* is very high (51,446 cases), so the full scale is not shown on the vertical axis. To fully represent the work the model is doing in correctly labeling *true negatives*, the vertical axis would be many times higher than is shown in *Figure 3*, and the comparatively small cluster of cases at the bottom of the graph with fine-grain distinctions between *true positives*, *false negatives* and *false positives* would not be discernable.

In the data, we have two groups of records: high-cost users (top 10 percent) and low-cost users (bottom 90 percent). High-cost users are illustrated by the blue curve and low-cost users are shown by the red curve. The cut-off threshold for including cases in the high-cost group based on their probability score is 0.37; any case at or above this level is estimated to be a high-cost user.

However, as seen in *Figure 3*, the model generates some errors in making these estimates, which are labeled *false negatives* or *false positives*. *False negatives* refer to records on the blue curve to the left of the cut-off level. These individuals were actually in the highest-cost 10 percent but the probability score produced by the model was below the 0.37 threshold in

Figure 3: Distribution of Probability Results for High and Low Cost Cases
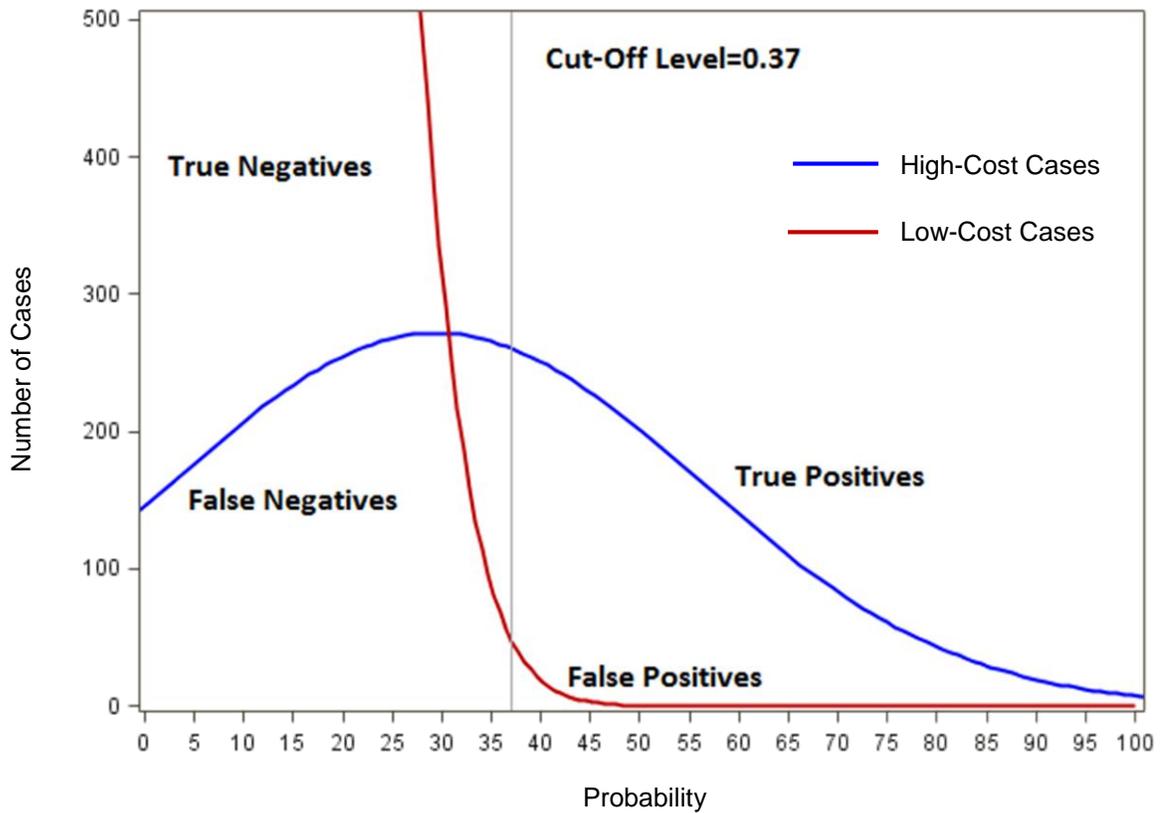
2012. Hence, they would not be selected for housing. *False positives* refer to records on the red curve to the right of the cut-off level. These individuals were not in the highest-cost 10 percent but scored higher than the 0.37 threshold and they would be incorrectly selected for housing.

The other two groups are correct estimates – *true positive* records on the blue curve to the right of the threshold and *true negative* records on the red curve to the left of the threshold. A model with a high predictive accuracy yields higher numbers of *true positives* and *true negatives* and lower numbers of *false negatives* and *false positives*.

At the selected cut-off level of 0.37, the Silicon Valley Triage Tool shows a low proportion of *false positives* relative to *true positives*. By selecting a lower threshold, it is possible to lower *false negatives* (and increase *true positives*) but the negative tradeoff is that there will be more *false positives*. The selection of the probability threshold for including cases in the high-cost group depends on the goal of the housing initiative and the financial capacity of the program to include lower-cost individuals. The nature of these trade-offs and costs associated with them are explored below.

True Positive Cases

The *true positive* group has average costs in the upper half of the high-cost range in all three years after the prediction. The average triage tool score for this group was 0.63. Costs declined after 2010, but are still over $54,000 in 2012. This reflects two factors, the first being *regression to the mean*, that is the tendency of extreme outcomes to be closer to the average when measured a second time. The second factor was declining public expenditures following the Great Recession. Over three-quarters of true positives stayed

True positive cases are more frequently diagnosed with schizophrenia than any other group.

Figure 4: Distinctive Attributes of True Positive Cases

Chart categories (top to bottom): Mental Disorder, Psychosis, Tri-Morbid, Homeless 2008-2010, Male, Skin Disease, Schizophrenia, Endocrine Disease, High-Med. Jail Security, 35-44 Years, Maximum Jail Security, 1+ Men. H. Inpat. Days, 4+ Emerg. Psych. Visits, 5+ Jail Men. Hlth. Days, 28+ Hosp. Inpat. Days, 5+ Hosp. Inpat. Admis., 12+ Emerg. Rm. Visits

X-axis: 0% to 100%

Legend: ■ True Positive ■ False Positive ■ False Negative ■ True Negative

in the high-cost group for over five years, while the remaining quarter shows post 2010 cost levels lower than pre 2010 costs significantly contributing to the decline in the overall cost average for this group from 2010 to 2012 that is shown in *Figure 2*.

Each of the four prediction groups has distinctive attributes that help explain its cost trajectory. The *true positive* group is differentiated by having acute problems and using expensive services more frequently than in any other group, as shown in *Figure 4*. 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 security 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 the most likely to be tri-morbid – diagnosed with a mental disorder, a chronic medical condition and to abuse drugs or alcohol. Demographically they are the most likely to be male and to be in the middle of their lives - 35 to 44 years old. And they are the 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.

The most distinctive features of the false negative group can be precursors to rapid escalation in public costs.

Figure 5: Distinctive Attributes of False Negative Cases

Legend: ■ False Negative ■ True Negative ■ True Positive ■ False Positive

Categories (top to bottom): Hospital inpatient 2010, 25-34 Years, Pregnancy Complication, 18-24 Years. X-axis: 0% to 60%.
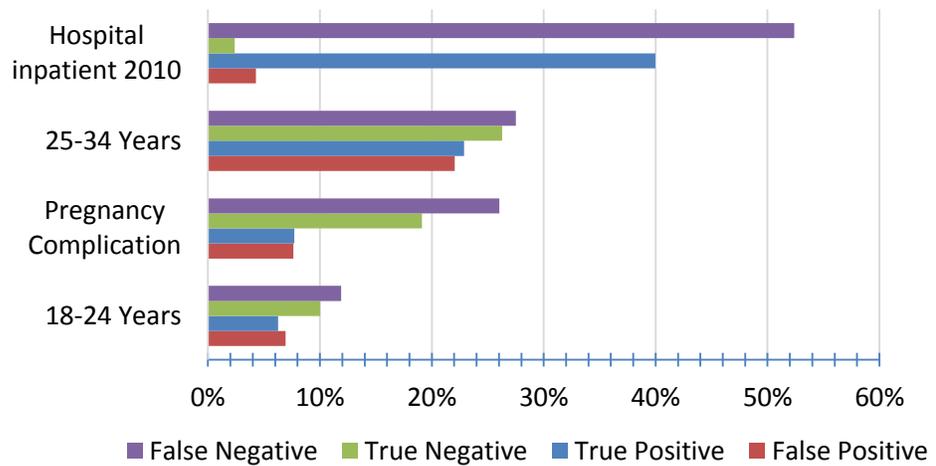
False Negative Cases

The *false negative* group had costs that increased 216 percent from 2009 to 2010, and then following that spike, drop as abruptly as they rose. Since the model predicts high-cost status in 2010 based on 2008 and 2009 service levels, even though the spike in 2010 was significant, low service levels in 2008 and 2009 did not generate high scores in 2010 for this group and they become *false negatives*. Moreover, as observed in *Figure 2*, the average cost of this group in 2010 is much lower than the *true positive* group. The average triage tool score for this group was 0.14. But based on the purpose of the model to identify individuals with continuing high costs, the preponderance of *false negatives* are in fact *true negatives* when we look at their costs over the three years following the prediction. The prevalence of one-time cost spikes in this group is borne out by the fact in 2012, only 22 percent of this group had cost over $25,000.

The most distinctive features of the *false negative* group are attributes that can be precursors to rapid escalation in public costs – youth and pregnancy, as shown in *Figure 5*. Individuals in this group are the most likely to be 18 to 34 years of age, and among women, to have a pregnancy with complications that result in high medical costs. This group had inpatient hospital costs in 2010, the prediction year, more frequently than any other group. However, we can infer from *Figure 2* that these hospitalizations were often for conditions that resulted in one-time cost spikes rather than continuing high costs.

> False positive group had a cost spike in the last year of data used to make the prediction.
>
> --------------------------

Figure 6: Distinctive Attributes of False Positive Cases

Legend: ■ False Positive ■ True Positive ■ False Negative ■ True Negative

Categories (top to bottom): Substance Abuse, Medium Jail Security, Minimum Jail Security, 45-54 Years. X-axis: 0% to 100%.
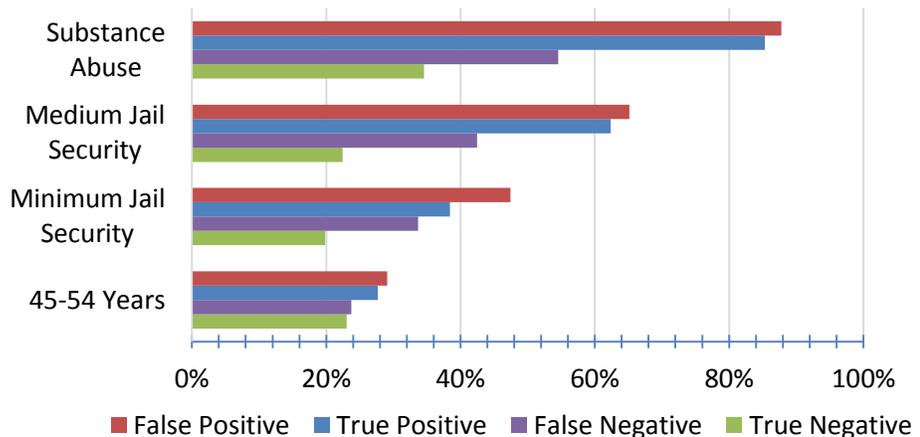
False Positive Cases

The *false positive* group also had a cost spike, but it occurred in 2009, the last year of data used to make the prediction, rather than in the year following the prediction. Average costs for this group decreased 76 percent from 2009 to 2010. The model did not anticipate this drop and included them in the high-cost category. The average triage tool score for this group was 0.53. However, after the drop, this group has increasing costs in 2011 and 2012. In 2012, 29 percent of this group had costs over $25,000, and the overall average cost for this group rose above the bottom threshold for the high-cost group. *False positives* fall into two cost groups of roughly equal size. Costs of the first group continued to stay at the low 2010 level for the next two years. However, costs of the second group rose steadily, reaching the level of *true positives* by 2012.

The most distinctive features of the *false positive* group are that individuals are most frequently older – over 45, incarcerated with low security risk classifications, and abusing drugs or alcohol, as shown in *Figure 6*. In addition, they are a close second to the true positive group in their rate of mental disorders, tri-morbidity, and being male, as profiled earlier in *Figure 4*. The cost spike in 2009 results in being classified as part of the high-cost group, but the problems that caused these costs abated in the prediction year, then began

Figure 7: Problems Present in False Positive Cases

Category	Percent
MEDICAL DIAGNOSIS	
Chronic Medical Condition	95%
Mental Disorder	88%
Neurotic Disorder	82%
Injury or Poisoning	72%
Psychosis	70%
Digestive Disease	60%
Musculoskeletal Disease	59%
Skin Disease	49%
Nervous System Disease	47%
Respiratory Disease	47%
Mental Health Inpat. 2008-09	39%
Infectious Disease	39%
Circulatory Disease	39%
Tenitourinary Disease	35%
Schizophrenia	33%
Hospital Inpatient 2008-2009	33%
Endocrine Disorder	31%
JUSTICE SYSTEM	
Jail 2008 or 2009	79%
Medium Jail Security	65%
Probation 2008-2009	39%
DISABILITIES	
Substance Abuse	88%
Dual Diagnosis	80%
Tri-Morbid	79%

to rise in subsequent years. In short, they have many of the attributes of the *true positive* group but their problems are not yet so severe as to result in the same level of ongoing intensive use of public services that characterizes the true positive group.

A profile of the false positive cases is shown in *Figure 7*. It shows that many problems including psychoses, serious health problems, substance abuse, and justice system involvement are prevalent in this group. This is a population with serious disabilities that needs housing, however in 2010, it was not yet a population with consistently high costs.

True Negative Cases

The *true negative* group contains over four-fifths of the cases and had continuing low costs, indicating that nearly all cases were correctly classified. The previous report, *Home Not Found*, showed that 83 percent of individuals who experience homelessness do not become persistently homeless, this includes 20 percent who are homeless only one month or less. These cases make up most of the *true negatives*. The average triage tool score for this group was 0.08. Only a small share of this group became long-term causalities of

The true negative group contains over four-fifths of the cases and had continuing low costs.

Figure 8: Distinctive Attributes of True Negative Cases
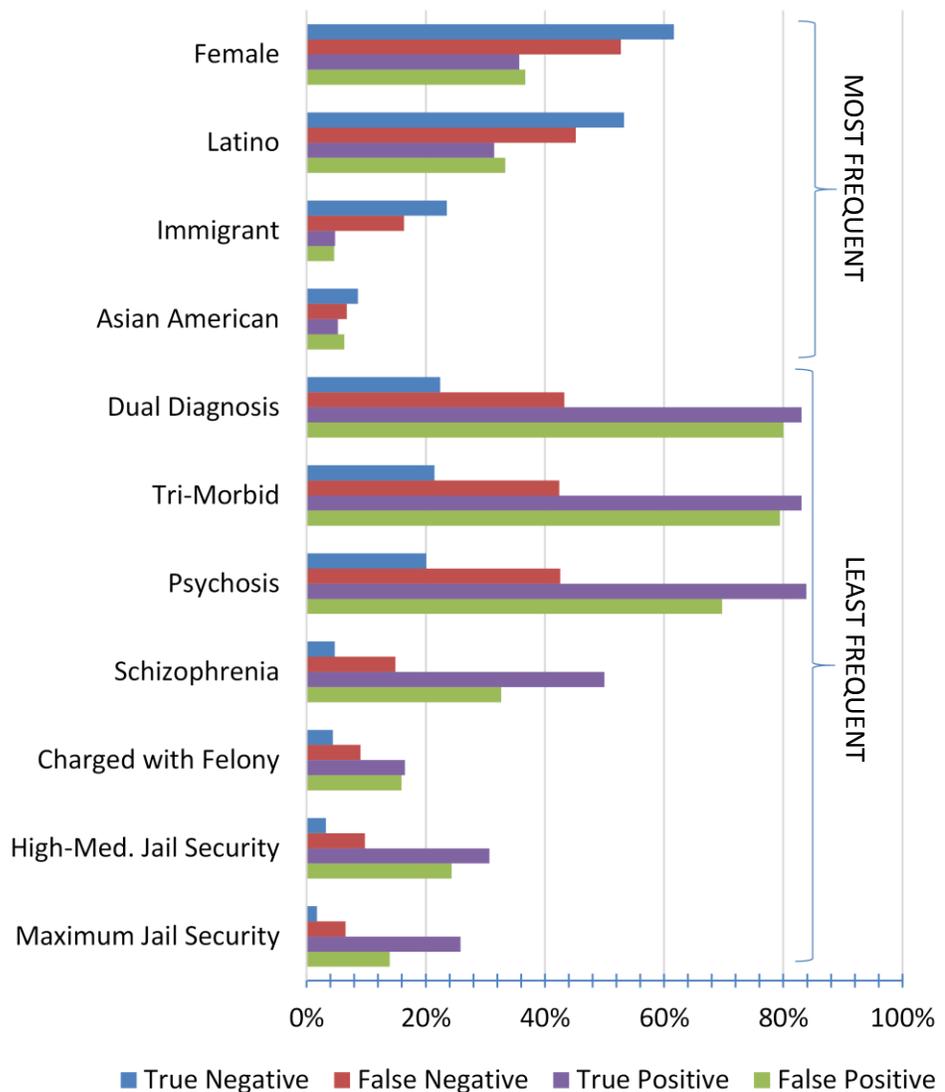
homelessness and had rising costs. Seven percent of this group had costs over $25,000 in 2012.

The *true negative* group is distinguished by two types of attributes – those that are more frequent than any other group and those that are less frequent than any other group, as shown in *Figure 8*. Individuals who are accurately classified as not being in the high-cost group are more frequently female, Latino, an immigrant, and Asian American. They are less likely than any other group to be dual diagnosed, tri-morbid, diagnosed with a psychosis, or within the different types of psychoses, to be schizophrenic. They are also least likely to have serious justice system involvement – to be charged with a felony or to have a maximum or high-medium jail security classification.

Methodology

Introduction

The purpose of triage tools is to identify persons experiencing homeless with a history of costly utilization of public services and so that they can be connected permanently affordable housing and cost-effective community-based health care and support services. The tool use administrative data to prioritize homeless adults with the highest needs and public costs. This approach requires some mechanism to identify or predict high-risk accurately before substantial preventable or avoidable costs have been incurred and the crisis status of homeless adults has deteriorated further. One such mechanism is a statistical predictive model, which is presented in this report.

In addition to the work done by Economic Roundtable in identifying high-cost homeless persons, a number of studies have proposed various methods to predict future high-cost health system users focusing on frequent users of hospitals and health clinics, each offering different models, predictor variables and types of data. These models were developed to identify high-risk patients at risk of readmission to 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). Some other studies estimated predictors of homelessness and developed methods for more efficient homelessness prevention services (Bryne et al. 2015). A recent study on New York City HomeBase 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).

Building upon previous research, we developed a predictive model to identify homeless individuals at risk of becoming high-cost public service users in Santa Clara County. The data sources, methods and results of this predictive modelling are presented below.

Data Sources

By collaborating in linking their client records, seven agencies in Santa Clara County provided the 38 pieces of information that are used by the triage tool to predict the probability that clients are in the high-cost group. In some instances such as with demographic information and medical diagnoses, the same information is aggregated from multiple agencies to ensure that it is complete. The information provided and the agencies contributing each type of information is shown in Table 1.

Methods

The purpose of the model presented in this report is to predict 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 years, using a predictive analytic modeling approach.

A predictive model is simply a mathematical function that maps the relationship between a set of input data variables and a response or target variable. Predictive analytics is an area of data mining that prepares and uses historical data to predict an unknown event of interest in the future applying analytical techniques such as regression models, decision trees or machine-learning algorithms. All these technical approaches provide a predictive score (probability) for each individual in order to determine priorities for intervention across large numbers of individuals.

By collaborating in linking their client records, seven agencies in Santa Clara County provided the 38 pieces of information that are used by the triage tool.

Table 1: Agencies Providing Data Used in the Triage Tool

Data Description	Medical Care	Mental Health	EMS	Drugs and Alcohol	Sheriff	Social Services	HMIS
DEMOGRAPHICS							
Age 18-45	X	X		X	X	X	X
Age 46-65	X	X		X	X	X	X
Age 65+	X	X		X	X	X	X
Female	X	X		X	X	X	X
CRIMINAL JUSTICE							
100+ days of probation in the last 2 years					X		
Arrested in last 2 years					X		
Jail booking in last 2 years					X		
Jail security classification of 3 or 4 this year					X		
Booking code 849 - arrested for inebriation, released					X		
Number of arrests this year					X		
Number of days in jail this year					X		
MEDICAL DIAGNOSES							
Diagnosed with chronic medical condition	X	X					
Adjustment reaction ICD-9 309 in last 2 years	X	X					
Heart disease ICD-9 401-429 in last 2 years	X						
Number medical encounters with Organ Failure in last 2 yrs.	X						
Schizophrenia ICD-9 295 in last 2 years	X	X					
Neoplasm (ICD-9 140 to 239) in last 2 years	X						
Diagnosis of ill-defined condition (ICD-9 799) in last 2 years	X						
High-cost ICD-9 in last 2 years (see list)	X	X					
HEALTH AND EMERGENCY SERVICES							
Emergency Medical Service (EMS) this year			X				
Emergency Medical Service (EMS) last year			X				
Two+ EMS encounters in last 2 years			X				
Hospital inpatient via ER or Emer. Psych. in last 2 years	X						
Outpatient Psych. Emer. Srv. or amb. surgery this year	X						
Number of hospital inpatient admissions this year	X						
Number of hospital inpatient days in last 2 years	X						
ER or medical clinic visit this year	X						
Number ER or clinic visits this year	X						
Eleven+ ER or clinic visits this year	X						
BEHAVIORAL HEALTH							
Number Mental Health outpatient days in the last 2 yrs.		X					
Two+ Mental Health outpatient visits in the last 2 years		X					
Number Mental Health inpatient admissions this year		X					
Two+ Mental Health inpatient admission in last 2 years		X					
Substance abuse - medical diagnosis, arrest or DADs	X	X		X	X		
Number of DADS encounters in the last 2 years					X		
HMIS AND SOCIAL SERVICES							
Chronic homeless flag in HMIS record							X
Public assistance benefits received in current year						X	
Two+ months of food stamps in the past 2 years						X	

EMS = Emergency Medical Services, DADs = Department of Alcohol and Drug Services, HMIS = Homeless Management Information System

We adopted a supervised training approach in this study because of the availability of data with known outcomes. In the supervised approach, the available sample records with known attributes and performances is referred to as the "training sample." The records in other samples, with known attributes but unknown performances, are referred to as "out of training" or "validation sample" records. Training is repeated until the model learns the mapping function between the given inputs and desired outcome. The model or algorithm classifies an observation by assigning a category to that piece of information.

Applying the predictive analytic methodology, we built a model that 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. Data from 2007 to

2009 was our training sample. We validated the model by applying the model to 2010 and 2011 records to predict high cost status in 2012. The details of model building and validation are presented later in this section.

One critical aspect of our methodology was to avoid the application of black-box analytics. Black-box refers to predictive modeling techniques used in machine-learning that do not explain their reasoning. Although extremely powerful, machine-learning techniques such as neural networks and support vector machines fall into this category. These algorithms are useful in classifying a high risk of churn for a particular customer or a high risk of fraud for a credit card transaction. However, these models do not explain why given types of information are used to make predictions. To make the triage tool understandable and credible to policy makers, service providers and the general public it is important to have reasonable explanations for why information being used to make predictions is relevant.

Since our predictive model is intended as a triage screening tool, we need to know which factors contribute to the final score or probability as well as their weights in order to design the tool. Moreover, the predictive model requires data elements from multiple public service domains ranging from hospitals to jails. Knowing the input factors used in the model is critical in building the logistics of data integration behind this model. Consequently, we only used regression models that are capable of explaining the classification or decision process rather than using machine-learning algorithms.

It is important to have reasonable explanations for why information being used to make predictions is relevant.

Model Variables

As presented earlier and elaborated in our earlier report (Economic Roundtable, 2015), we used an integrated database built by linking eleven administrative data sources. All these datasets include information on factors that may have an effect on our outcome of interest—becoming 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 current and previous years (e.g., number of clinic or emergency room visits, number of hospitalizations, number of arrests) as well as the cost of services.

First, we generated a binary target variable to flag whether or not homeless persons were top 10 percent high-cost users in 2009 (training cohort) and 2012 (validation cohort). In order to identify high cost status we summed costs across all service types and ranked them separately for these two cohorts.

The next step was to identify any potential variables that would have an effect on becoming a high-cost user. Since each data source has many variables, this step required a laborious process to prepare all potential variables and select the first set of candidates. In this step we selected relevant diagnostic codes and service factors combining our experience in Los Angeles and a review of the literature with statistical tests of association––applying *chi-square* and *t-tests* to assess whether any of these factors contribute in separating high-cost users from others. This step generated the first iteration of variable selection after eliminating redundant and irrelevant factors. We tested hundreds of potential variables available from agency databases including demographics, clinical and service utilization variables and identified almost 100 variables for model selection. The list of these variables is shown in *Appendix Table A-1.*

After selecting our initial set of variables, we prepared the data by transforming variables in the pre-processing stage. Data pre-processing augments the predictive power of variables by transforming and preparing them for model development. Continuous fields may be binned (such as the age category which is modified into 3 groups—18 to 45, 45 to 65, and

65 or older. Some categorical variables were clustered such as ethnicity and diagnostic codes. A majority of the variables were transformed into binary (1 or 0) variables, for example, for whether or not an individual had a given medical diagnosis. These variables equal 1 if a condition exists (such as diabetes) and 0 if the condition does not exist. All these binary variables were generated for the current and previous years. We generated many count variables that show the number of occurrences of a variable such as hospital or emergency room visits, arrests, and days on probation. Finally some count variables were transformed into binary variables such as under or over 100 days of probation. All count variables were also generated for the current and previous years.

Model Development

After completing the pre-processing stage, we had over 250 input variables to be trained in our predictive models. In the next step, we built several models to test their performance in predicting high-cost users. We used the SAS Enterprise Miner platform to develop and assess predictive models (See Sarma 2013; SAS 2013).

As noted earlier, we avoided using black-box analytics and selected regression techniques that identify the variables contributing to our prediction. Regression models are the mainstay of predictive analytics. The focus lies on establishing a mathematical equation as a model to represent the interactions between the different factors that affect an outcome, such as being a high-cost user of public services. We used a *decision-tree* model, a *least-angle regression* (LARS) model and a *logistic regression* model applying the backward variable selection methods.

A decision tree represents an algorithm using a tree-like model of decisions and their outcomes. Each node in the tree represents a "test" on an attribute (e.g. whether there was an arrest or not), each branch represents the outcome of the test and each leaf node represents a classification—high cost or not (See, De Ville and Neville, 2013). LARS is an algorithm for fitting linear regression models to high-dimension data with its own selection method (Efron, et.al. 2004). Logistic regression is one of the most commonly used techniques in predictive analytics. It is a technique in which unknown values of a discrete variable (high-cost user or not) are predicted based on known values of multiple variables (See Allison, 2012). We used backward selection technique with our logistic regression model. Under this approach one starts with fitting a model with all the variables of interest. Then the least significant variable is dropped, so long as it is not significant at our chosen critical level—5 percent. We continue by successively re-fitting reduced models and applying the same rule until all remaining variables are statistically significant.

The population of homeless persons was 57,259. We only included individuals with at least one linked record to an agency during our study window between 2007 and 2012. Since there were record linkage problems with databases of some agencies, we selected only those individuals with linked records (See Economic Roundtable, 2015). As mentioned earlier, the training dataset included 2007 to 2009 records and the validation dataset included 2010 to 2012 records.

We used each model to predict the status of each person in the dataset as a high cost user in the next year. Model performances were compared using SAS Enterprise Miner. Among the three models we trained, logistic regression performed the best and was selected as the champion model. In the final phase we fine-tuned this model by 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 can be aggregated under one variable without sacrificing the model performance. The results of this final model are presented in the next section.

Results

The final model had 38 variables with main effects and 11 variables with interactions. The descriptive values of model variables are shown in *Appendix Table A-2*. Performance of the model was evaluated using *C-statistic* to assess the predictive ability of the model. Significance of the parameter estimates (*p-values*) and *odds ratios* were evaluated as well. The *odds-ratios* of the final model are presented in the *Appendix, Table A-3*.

As shown in *Appendix Table A-2*, 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.

We tested 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, we used two other factors which are the aggregations of chronic medical conditions and high-cost ICD-9 codes as presented in *Appendix Table A-4.* The high-cost homeless group shows much higher rates of encounters with these diagnoses while overall averages vary between 6 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 20 percent 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. In contrast, there is little difference in social services participation rates.

Adjusted odds ratios are presented in *Appendix Table A-3*. The results reflect the differences we observe from descriptive comparisons. Logistic regression models generate odds ratios that are used to assess the likelihood of a particular outcome (being a high-cost top 10 percent person in this study) if a certain factor (one of the model variables) is present. It is a relative measure showing how likely a person with a certain attribute (say female) is to experience the outcome (high-cost person) relative to another person without the attribute (male). In this way, we capture the strength of relationship between the factor (say gender) and the outcome. Adjusted odds ratios are generated after controlling for all other variables in the model, which means holding all other factors constant.

Odds ratios for binary variables (for example, arrested or not) are in general higher than the odds ratios for continuous variables (for example, days in jail) and are interpreted differently. *Appendix Table A-3* shows 95 percent confidence limits so that we are 95 percent certain that the true population odds ratio falls in this range.

The high-cost group shows higher rates of engagement with health and emergency services.

_____

Being arrested in the last two years, higher jail security and substance abuse are among the strongest predictors of becoming a high-cost homeless resident.

For example, the odds ratios show that persons who have been arrested in the past 2 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 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 persons with adjusted ratios in the range of 1.05 and 1.28. One exception is receiving two or more months of food stamp payments, which has an odds ratio of .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 DADS encounters) to 1.16 (number of hospital admissions), and all increase the likelihood of becoming a high-cost homeless person.

The model achieved a very strong *C-statistic*: .813. *C-statistic* is the probability that predicting the outcome is better than chance. It is used to compare the goodness of fit of logistic regression models. Values for this measure range from 0.5 to 1. A value of 0.5 indicates that the model is no better than chance at making a prediction of membership in a group; a value of 1 indicates that the model perfectly identifies those within a group and those not. Models are typically considered reasonable when the *C-statistic* is higher than 0.7 and strong when *C* exceeds 0.8 (Hosmer and Lemeshow 2000). Another widely used measure of model performance is the *Brier score* which is the mean squared difference between the predicted probability and the actual outcome. The lower the *Brier score* is for a set of predictors, the better the classification performance of the model (perfect score is 0). The *Brier score* for the model is 0.059, which is a very strong statistic.

Overall the model predicts high-cost homeless persons with a good fit. However, we needed to assess real model performance using validation tools to evaluate the out-of-sample prediction power, that is, prediction power for cases other than those used to develop the model. The next section presents the validation results.

Validation and Assessment

Statistical Tests

Validation of the model on the 2007-2009 cohort used to "train" the model was done using the 2010-2012 validation cohort to assess the out-of-sample predictive power of the model. We often observe strong predictive power based on in-sample performance if the model over-fits the data. In these 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. We used *sensitivity*, s*pecificity*, *positive predictive value* (*PPV*), and *accuracy* as well as the area under the receiver operating characteristics (*ROC*) *curve* to assess the out-of-sample model performance (See Gonen, 2007 for *ROC* analysis for predictive models). All these values are presented for different scenarios-top 1 percent, 5 percent, 10 percent as well as top 1,000 homeless persons with the highest risk of becoming a high-cost service user.

Table 2 presents *sensitivity*, *specificity*, *PPV*, and *accuracy* statistics for different cut-off points for the validation (out-of-sample) cohort. 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-specificty) 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. If *PPV* equals 1 this means that the model identifies all high-cost persons correctly—with no *false positives*. The higher the *false positives*, the lower the *PPV*. Finally, the *accuracy* statistic measures the proportion of *true positives* and *true negatives* out of all persons. If there are no *false positives* and *false negatives* (perfect model), the *accuracy* value would be 1.

Table 2: 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

TP – True Positive FP – False Positive TN – True Negative FN – False Negative N – Number in Population

If the top 5 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 5 percent are also very high. If we follow a subset within the Top 5 percent, the 1,000 cases (1.75

Figure 9: Lift Chart

Homeless Persons Ranked by Probability of Being in High-Cost Group
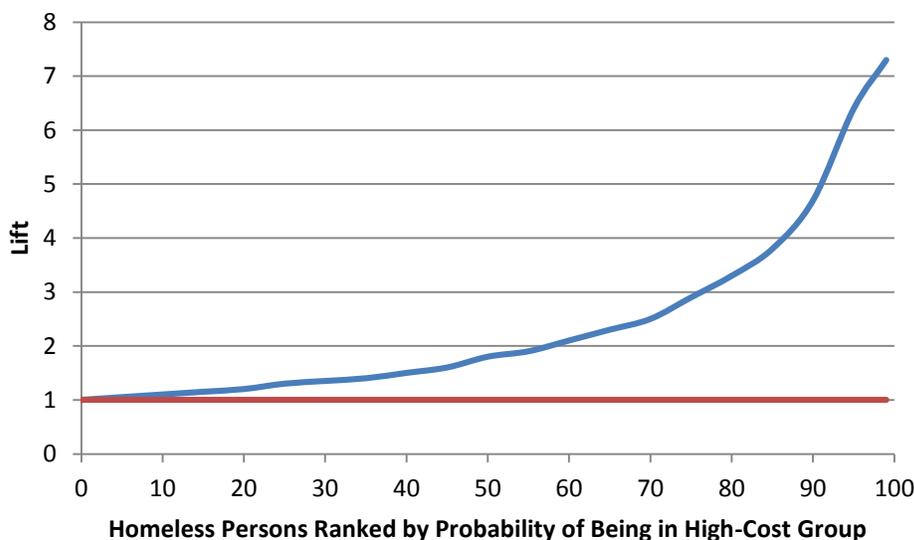
percent of all cases) with the highest probability scores of being in the high-cost group, we see even more accurate prediction outcomes. The model achieves a *PPV* result of 67 percent, meaning that out of 1,000 persons that model identified as high-cost persons, two-thirds are *true positives* and the remaining one-third are *false positives*. *PPV* is an important measure for assessing the cost-effectiveness of the model.

If we follow 1,000 randomly chosen persons, the *PPV* value would be only 10 percent. This means that, a random selection—without using any knowledge or model would yield only 10 percent *true positives*. The remaining 90 percent would be *false positives*. When we compare this number to the model *PPV* for 1,000 persons, we get a ratio of 6.7 which shows that model is performing approximately 7 times better than random selection.

Lift

A similar approach is known as *lift* charts in the literature. *Lift* is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model for all thresholds. *Figure 9* 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 5 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 1 or no *lift*. At slightly lower thresholds, such as the top 10 percent, *lift* drops to 4.7 because in order to classify more *true positives* we have to increase the number of *false positives*, but the number of *false positives* decreases as the probability of being in the high-cost group increases.

ROC Curve

Another way of assessing the predictive power of a model is the area under the *ROC curve*, which 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 10*.

Figure 10: ROC Curve
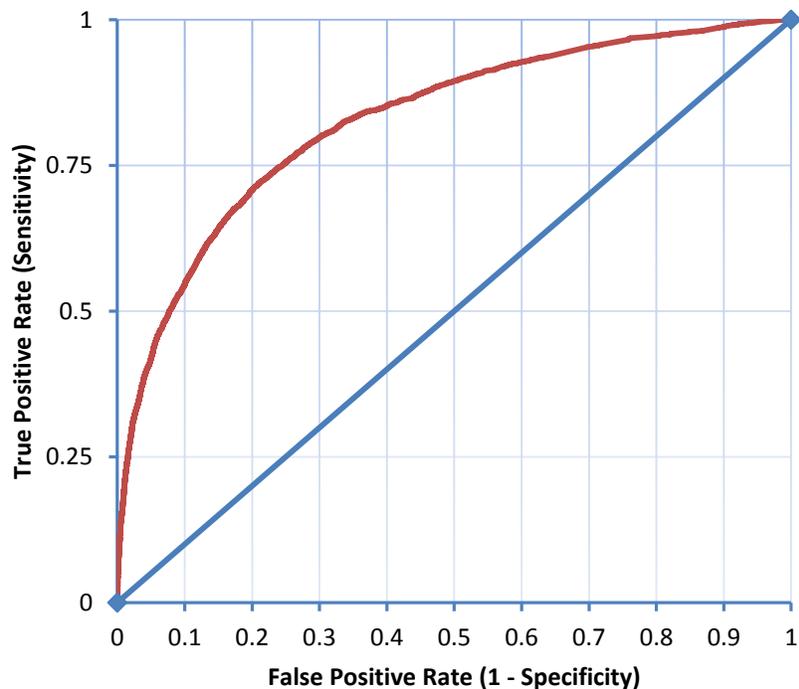
The closer the curve follows the vertical axis and then the top border, the more accurate the model. Conversely, the closer the curve comes to the 45-degree diagonal, the less accurate the model is. The *area under the curve* (*AUC*) measures the accuracy of the model where 1 represents a perfect model and 0.5 (same as the diagonal line) shows a useless model. 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 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 refers to poor model performance.

Since the model provides a probability score in the range of 0 and 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 that 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. This trade-off can be also presented by plotting proportions of *true positives* and *false positives* in the total population. These trade-offs are illustrated below in *Figures 11* and *13*.

In general, a low cut-off threshold yields a disproportionate number of *false positives* relative to the number of *true positives* and small numbers of *false negatives*. In contrast, a high cut-off threshold creates a disproportionate number of *false negatives* while avoiding large numbers of *false positives*. *False negatives* refer to high-cost homeless persons that model misses at a given threshold. We correctly identify high-cost users at the cost of missing many of them because there will not be many observations with very high scores.

The optimal cut-off is not simply an empirical decision. It depends on the availability of housing and the trade-offs between costs and savings that accrue based on who receives the housing.

True vs. False Positive Trade-Offs

The proportion of *true positives* and *false positives* in the total population are compared at different cut-off thresholds for inclusion in the high-cost group in *Figure 11*. The rate of *false positives* is very high at the lowest cut-off level and decreases as cut-off values increases. On the other hand, the rate of *true positives* starts low and increases at higher cut-

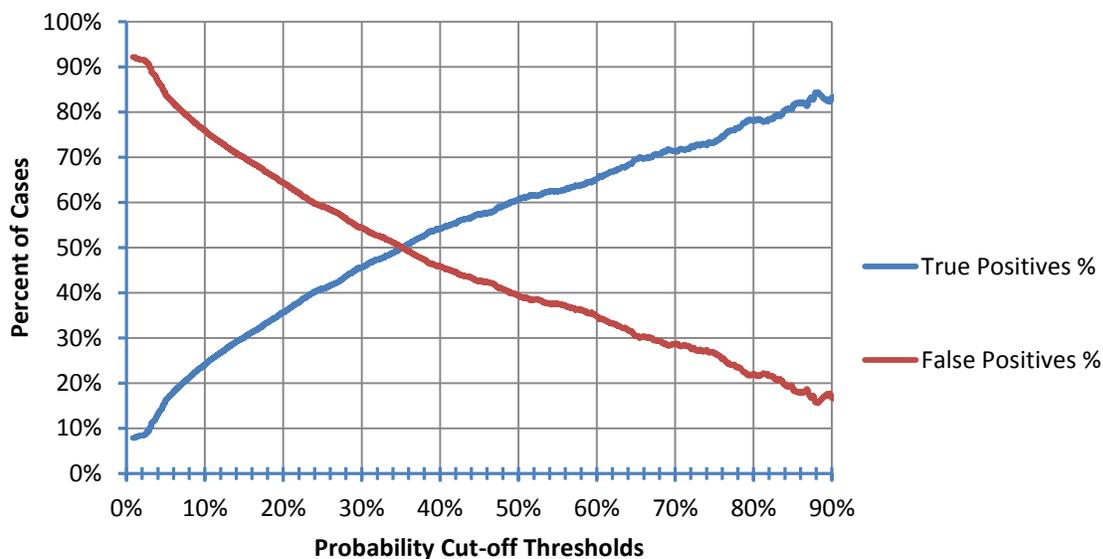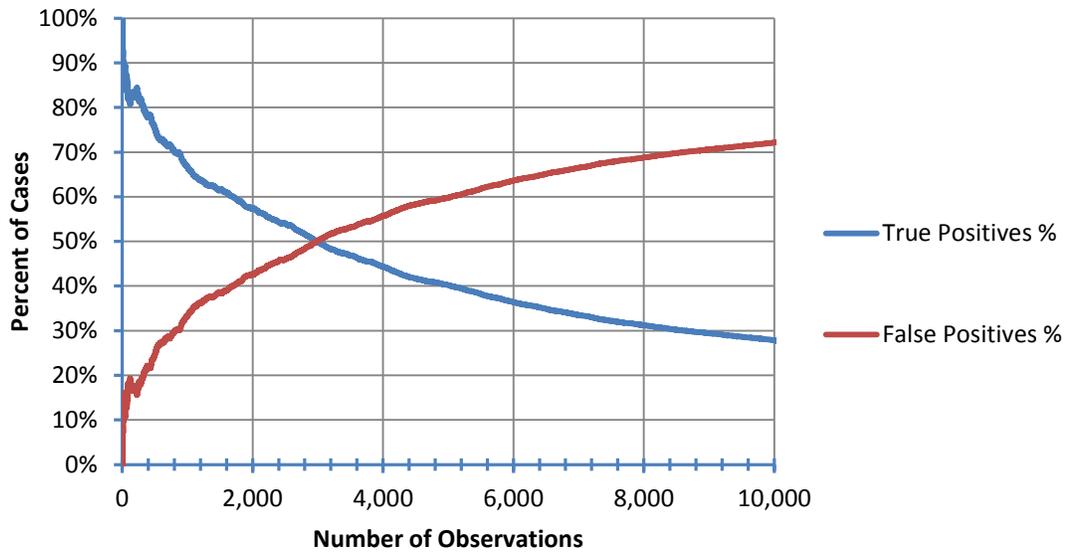**Figure 11: True and False Positive Percentage by Cut-off Point**

Figure 12: True and False Positive Percentage by Number of Observations

off values. They cross each other at the 33 percent cut-off level, suggesting this level offers equal trade-offs. At cut-off levels lower than 33 percent, we get more *false positives* for each *true positive* we capture. However, the optimal cut-off is not simply an empirical decision. In the context of permanent supportive housing, it depends on the availability of housing and the trade-offs between costs and savings that accrue based on who receives the housing. This is discussed in the next section. When we select a specific threshold, we have to show that the savings accrued from *true positives* (high-cost homeless persons) will be sufficiently high to offset shortfalls in post-housing cost savings for *false positives* (lower cost homeless persons). In addition, the threshold should identify a high proportion of high-cost homeless persons who represent the most vulnerable group with the highest level of service needs.

The distribution of true positives and false negatives when using the triage tool to identify the target population of almost 4,500 high-cost homeless persons is shown in *Figure 12*. Given the large number of high-cost individuals identified in Santa Clara County, the most effective use of the tool is to screen a database of linked records and prioritize access to permanent supportive housing based on individuals' probability of being in the high-cost group. If triage tool screening is used to offer housing to the 1,000 people identified as having the highest probability of being in the high-cost group, the chart confirms that approximately two-thirds of this population will be *true positives*, that is, high-cost service users. If the cut-off threshold is lowered to include the 3,000 people with the highest probability of being in the high-cost group, the ratio of *true positives* to *false positives* will drop to 50 percent.

In practice, it is important to note that the county does not have 3,000 or even 1,000 housing units to offer at a single time. New housing is being developed and made available over time. Until more effective preventative programs are initiated, it is likely that more individuals will continue to enter the high-cost group, so there is likely to be continuing justification for retaining a high cut-off threshold.

We have shown that at a given threshold that corresponds with selecting a target of a given size to be offered housing, there are trade-offs among *true positives* who are the correctly identified high cost persons and *false positives* who are incorrectly identified low cost persons. The distribution of *true positives* and *false positives* for the top 1,000 homeless persons is shown in *Figure 13*. At each score threshold, most of the homeless persons were identified correctly as high service users. However, a significant (approximately one-third) proportion of them were *false positives*, particularly in the range between the scores of 62

> The most effective use of the tool is to screen a database of linked records and prioritize access to permanent supportive housing based on individuals' probability of being in the high-cost group.

Figure 13: Distribution of True and False Positives for Top 1,000 Scores in 2010
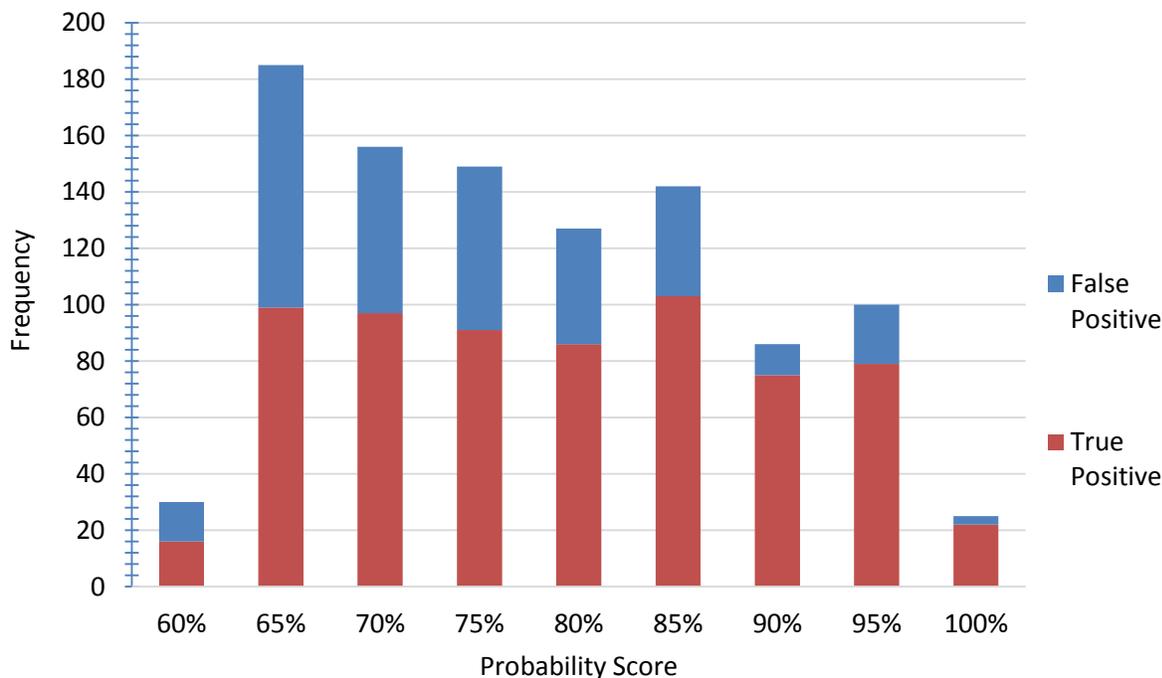
and 80 percent. Next, we assess the changing status of *false positives over* time in order to understand this group better

False Positives over Time

To assess *false positives*, we ran the Silicon Valley Triage Tool using 2008 and 2009 data to predict 2010 outcomes, which were then compared to actual 2010 costs to identify *true positives* and *false positives*. Using the scored 2010 data we can track cost outcomes for *true positives* and *false positives* for two additional years – 2011 and 2012. Cost outcomes for 2010 through 2012, broken out by probability score are shown for the 1,000 cases with the highest probability scores in *Table 3*. Rows in this table show results of the model's predictions in 2010. Columns in the table show the cost group that cases were in after 2010, based on average annual costs in 2011 to 2012. The breakout of cases in different rows is:

- *False positives in the 0-79 cost percentile range in 2010* – far below the high-cost group.
- *False positives in the 80-89 cost percentile range in 2010* – just below the high-cost group.
- *False positives because of no data in 2010* – cases with costs in 2008 and 2009 but not 2010.
- *True positives* - cases that were correctly predicted to be in the high-cost group in 2010.

Similarly, the column categories that show cost groups in 2011 and 2012 are broken out as follows:

- *Low-cost* in the 0-79 cost percentile range in 2011-2012.
- *Medium cost* - in the 80-89 cost percentile range in 2011-2012.
- *No cost data* – missing from records in 2011-2012.
- *High cost* – in the 90-99 cost percentile range in 2011-2012.

Table 3. True vs. False Positive Distributions of 2010 Scores—Top 1,000

Prediction Results in 2010 based on 2008-2009 Data	Cost Group in 2011 and 2012				
	Low-Cost (Percentiles 0-79)	Medium Cost (Percentiles 80-89)	No Data Missing from County 2011-2012	High-Cost (Percentiles 90-99)	Total 2011-2012
False Positives 2010 0-79 Cost Percentiles	37	10	14	42	103
False Positives 2010 80-89 Cost Percentiles	48	36	4	85	173
False Positives Missing in 2010	28	2	12	6	48
True Positives 2010	78	57	13	528	676
Total 2010	191	105	43	661	1,000

This breakout of changes in cost from 2010 to 2012, based on predictions using 2008–2009 data enables us to see the relationship between model probability scores and actual multi-year outcomes. We would expect that some *true positives* in 2010 falling out of the high-cost group in later years, while some *false positives* would move into the high-cost group in later years. This expectation turns out to be correct. There were 676 high-cost cases in 2010 and 661 high-cost cases in 2011-2012. This demonstrates that the proportion of two-thirds of the top 1,000 being in the high-cost group in a given year holds true over a multi-year period.

It is difficult to differentiate the *medium cost false positives* from the *true positives* for many records. The *false positives* may be in the 88th or 89th percentile, showing only slightly less need and service use in 2008 and 2009 data when predicting 2010 outcomes. Many of them turn out to be high users in future years. We observe in *Table 3* that out of 324 *false positives* in 2010, more than half (173) of them are in the *medium cost false positive* group. Almost half (85) of this *medium cost* group became high users during the next 2 years, while almost a quarter (36) stayed in the *medium cost* group, and a little over a quarter (48) moved into the *low cost* group.

The least successful predictions are the *low-cost false positives*. This group represents persons with very high costs in 2009 (and sometimes 2008 as well) but who then used very low levels of services during the next 3 years, ending up in the 0-79th percentile cost range. The model predicted them to be in the top 10 percent in 2010 because of their cost spike in 2008 to 2009. However, they make up only a small proportion of all predictions. *Table 3* shows that, in addition to 37 persons who remained *low-cost false positives* for all three years, there were another 88 persons who joined this low-cost group after 2010. This makes 125 persons, representing only 12.5 percent of the top 1,000 persons.

This analysis shows that when we follow cost outcomes for the top 1,000 scored persons, the model correctly identifies two-thirds of homeless persons with high cost utilization in the next year. Over the three years following the prediction data, 12.5 percent of the high scores are *true false positives*, representing homeless persons with a spike in their cost utilization in 2008-2009 that disappeared in the following three years. Another 15 percent have medium but not-quite-high costs that are in the 80-89th cost percentile range. Finally, less than 5 percent of the population disappears in the next year with no service use for a variety of reasons such as moving out of the county, dying or being institutionalized.

Cost Savings and Business Case

Cost Distribution among Residents Experiencing Homelessness

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. 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 permanent supportive housing to the targeted population. The following analysis explores the cost efficiency of providing permanent supportive housing to targeted high-cost homeless persons.

The results of the model show that the distribution of actual costs by probability score is skewed very heavily to the right. The distribution of actual costs in 2010 for 1,000

Figure 14: Cost Distribution for Top 1,000 Probability Scores in 2010
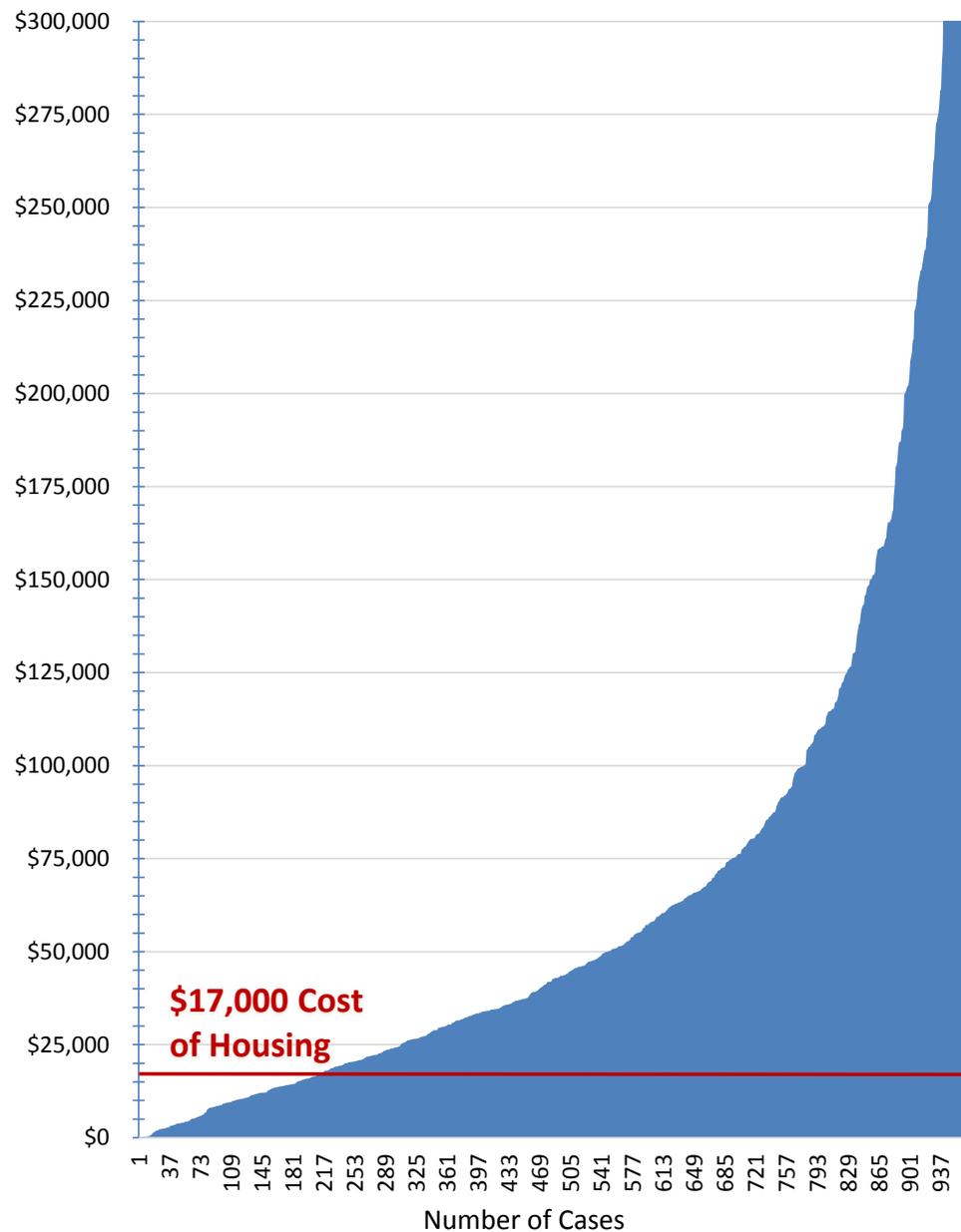
individuals with the highest probability scores for being in the high-cost group based on 2008-2009 data is shown in *Figure 14*. The probability range for this group was 0.66 to 0.99. The average cost in 2010 for these 971 individuals (29 had no cost data in 2010) was $70,089. Only 216 individuals had public costs that were less than the estimated $17,000 annual cost to provide housing. Twenty-six individuals had costs over $300,000, including three with costs over $1,000,000. The highest cost was over $1.5 million for a 32 year old man diagnosed with schizophrenia and a musculoskeletal disorder who was a hospital inpatient for 322 days and spend another 10 days in jail mental health incarceration during the year.

Business Case Analysis

Using five years of actual cost data, from 2008 through 2012, we are able use the first two years of data to produce triage tool probability scores for the likelihood of each individual being in the highest-cost group, and then track the accuracy and financial outcomes of these predictions over the following three years. An overview of these outcomes was shown earlier in *Figure 4*. Two prominent findings from this analysis are, first, that public costs for individuals experiencing homelessness vary significantly from one year to the next and, second, ranking individuals based on triage tool probabilities is effective for identifying the highest-cost population over a multi-year period.

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 *false positives* with no expected cost savings. Our estimation of net savings at different cut-off levels estimates 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.

In this analysis we look 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 10 percent of our test population with complete record linkage data (or 5 percent of the total, multi-year population experiencing homelessness) as high-cost users. The 0.53 cut-off level identifies the top 1,000 high-probability service users in our test population.

We assume that the annual cost of permanent supportive housing is $17,000 per person per year, based on rent subsidy and supportive service costs in Los Angeles. Finally, we assume that the post-housing reduction in service costs is 68 percent for homeless persons in the 10th decile based on a recent study from Los Angeles (Economic Roundtable, 2009). Other studies from Los Angeles and Charlotte, North Carolina confirm that expected service cost reductions for homeless persons in permanent supportive housing are in the range of 60 to 80 percent (Toros and Stevens, 2012, Thomas, et.al. 2014).

We also assume that there will not be any anticipated 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).

Since we have actual cost data each of the three post-prediction years, 2010-2012, we are able to compute actual cost savings for each year. Post-housing costs savings are calculated as 68 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 post-housing costs from homeless costs for the year. We used 2014 prices for the analysis and estimated cost savings for those cases with service utilization during all five years - 2008 through 2012.

We estimate saving of over $22 million in 2011 if the top 5 percent of high cost service users were housed permanently with supportive services.

In our analysis we estimated cost differences for six probability-cost groups, which all 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, *true positives* in 2010 may become high or low cost service users in the future. We evaluated the long-term cost status of individuals based on their actual cost rankings in 2011 and 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*. Finally, if a score was below the selected cut-off and 2010 costs were in the top decile, the record is a *false negative*. *False positives* and *negatives* may become high or low cost service users in the future, which we tested by observing actual costs in 2011 and 2012, identifying cases that moved into the *true positive* cost category. Since we use actual costs in 2011 and 2012, *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.

Costs and Savings in 2008 to 2010

Costs in 2008 to 2010 for individuals with probability scores of 0.37 or higher (*true* and *false positives*), as well as for individuals with 2010 costs in the 10th decile whose but tool scores below 0.37 (*false negatives*), each broken out into subgroups of low- and high-cost users, are shown in *Table 4*. Data from 2008 and 2009 was used to predict whether individuals would have high costs in 2010. The *true positives* had steadily increasing costs in

Table 4: Costs and Savings, 2008–2010, at the Probability Cut-off Level of 0.37

Probability-Cost Group	DATA USED FOR TRIAGE TOOL SCORE		YEAR PREDICTED	POST-HOUSING COSTS AND SAVINGS				
	Average 2008 Cost per Person	Average 2009 Cost per Person	Average 2010 Cost per Person Unhoused	2010 Cost Savings per Person Housed (68% for 10th decile persons)	2010 Adjusted Cost per Person Housed ($17,000 added for housing cost)	2010 Net Savings per Person Housed	2010 Total Savings for Group when Housed	Number of Cases
PERSONS WITH TRIAGE TOOL PROBABILTY OF 0.37 OR HIGHER								
True Positives-Low Cost Users in 2011 to 2012	$51,574	$66,462	$90,989	$61,873	$46,117	$44,873	$11,442,545	255
True Positives-High Cost Users in 2011 to 2012	$70,235	$83,628	$93,196	$63,373	$46,823	$46,373	$35,614,729	768
False Positives-Low Cost Users in 2011 to 2012	$33,957	$49,461	$11,444	$0	$28,444	($17,000)	($8,823,000)	519
False Positives-High Cost Users in 2011 to 2012	$49,583	$51,015	$13,029	$0	$30,029	($17,000)	($5,899,000)	347
AVERAGE COST PER PERSON WITH PROBABILITY OF 0.37 OR HIGHER	$53,955	$65,933	$55,711	$34,118	$38,593	$17,118		
TOTAL FOR PERSONS WITH PROBABILITY OF 0.37 OR HIGHER	$101,920,643	$124,546,938	$105,237,525	$64,448,274	$72,902,251		$32,335,274	1,889
PERSONS IN THE 10TH COST DECILE IN 2010 WITH TRIAGE TOOL PROBABILITY BELOW 0.37								
False Negatives-Low Cost Users in 2011 to 2012	$12,619	$14,349	$50,723	$34,492	$33,231	$17,492	$24,977,881	1,428
False Negatives-High Cost Users in 2011-2012	$20,671	$25,608	$72,621	$49,382	$40,239	$32,382	$29,241,343	903
AVERAGE COST PER 10TH DECILE PERSON WITH PROBABILITY BELOW 0.37	$15,738	$18,710	$59,206	$40,260	$35,946	$23,260		
TOTAL FOR 10TH DECILE PERSONS WITH PROBABILITIES BELOW 0.37	$36,685,503	$43,614,128	$138,009,153	$93,846,224	$83,789,929		$54,219,224	2,331

each year from 2008 through 2010. The *false positives* had costs that increased from 2008 to 2009, but then decreased in 2010. The *false negatives* had low costs in 2008 and 2009, but high costs in 2010. If all 1,889 individuals with probabilities of 0.37 or higher had been housed in 2010, the average reduction in net public costs would have been $17,118 per person or over $32 million for the entire group.

Costs in 2008 to 2010 for individuals with probability scores of 0.53 or higher, as well as for individuals with 2010 costs in the 10th decile whose but tool scores below 0.53 are shown in *Table 5*. Because of the higher probability cut-off level, the group is smaller, 1,000 people, and has higher costs. The pattern of changes in annual costs from 2008-2010 is the same as in *Table 4*, but the costs are higher. Homeless costs in 2010 for individuals with probabilities of 0.53 or higher averaged $69,565. If this group had been housed, net post–housing costs would have averaged $27,242 per person or over $27 million for the group.

Table 5: Costs and Savings, 2008–2010, at the Probability Cut-off Level of 0.53

Probability-Cost Group	DATA USED FOR TRIAGE TOOL SCORE		YEAR PREDICTED	POST-HOUSING COSTS AND SAVINGS				
	Average 2008 Cost per Person	Average 2009 Cost per Person	Average 2010 Cost per Person Unhoused	2010 Cost Savings per Person Housed (68% for 10th decile persons)	2010 Adjusted Cost per Person Housed ($17,000 added for housing cost)	2010 Net Savings per Person Housed	2010 Total Savings for Group when Housed	Number of Cases
PERSONS WITH TRIAGE TOOL PROBABILTY OF 0.53 OR HIGHER								
True Positives-Low Cost Users in 2011 to 2012	$74,953	$82,150	$111,580	$75,874	$52,706	$58,874	$7,182,682	122
True Positives-High Cost Users in 2011 to 2012	$82,104	$99,338	$96,892	$65,886	$48,005	$48,886	$25,958,717	531
False Positives-Low Cost Users in 2011 to 2012	$38,866	$67,557	$12,427	$0	$29,427	($17,000)	($3,094,000)	182
False Positives-High Cost Users in 2011 to 2012	$64,066	$68,324	$13,579	$0	$30,579	($17,000)	($2,805,000)	165
AVERAGE COST PER PERSON WITH PROBABILITY OF 0.53 OR HIGHER	$70,386	$86,340	$69,565	$44,242	$42,322	$27,242		
TOTAL FOR PERSONS WITH PROBABILITY OF 0.53 OR HIGHER	$70,385,762	$86,339,668	$69,564,619	$44,242,399	$42,322,220		$27,242,399	1,000
PERSONS IN THE 10TH COST DECILE IN 2010 WITH TRIAGE TOOL PROBABILITY BELOW 0.53								
False Negatives-Low Cost Users in 2011 to 2012	$14,143	$17,594	$52,544	$35,730	$33,814	$18,730	$29,237,744	1,561
False Negatives-High Cost Users in 2011-2012	$25,483	$30,370	$75,177	$51,120	$41,057	$34,120	$38,897,355	1,140
AVERAGE COST PER 10TH DECILE PERSON WITH PROBABILITY BELOW 0.53	$18,930	$22,986	$62,097	$42,226	$36,871	$25,226		
TOTAL FOR 10TH DECILE PERSONS WITH PROBABILITIES BELOW 0.53	$51,128,627	$62,085,367	$167,723,675	$114,052,099	$99,588,576		$68,135,099	2,701

Cost Savings in 2011

Actual 2011 costs for individuals with probability scores of 0.37 or higher, as well as for individuals with 2010 costs in the 10th decile whose but tool scores below 0.37 are shown in *Table 6*. 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.

Based on our assumptions, we estimate saving of over $22 million in 2011 if the 5 percent with the highest probability of being high cost service users were housed permanently with supportive services. 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 5 percent of the population identified by the triage tool as having the greatest probability of high future costs.

Table 6: Cost Savings for 2011 at the Cut-off Level of 0.37

Probability-Cost Group	YEAR PREDICTED 2010 Average Cost per Person if Unhoused	1 YEAR AFTER PREDICTION 2011 Average Cost per Person if Unhoused	POST-HOUSING COSTS AND SAVINGS 2011 Cost Savings per Person Housed (68% for 10th decile persons)	2011 Adjusted Cost per Person Housed ($17,000 added for housing cost)	2011 Net Savings per Person Housed	2011 Total Savings for Group	Number of Cases
PERSONS WITH TRIAGE TOOL PROBABILTY OF 0.37 OR HIGHER							
True Positives-Low Cost Users in 2011 to 2012	$90,989	$10,932	$0	$27,932	($17,000)	($4,335,000)	255
True Positives-High Cost Users in 2011 to 2012	$93,196	$83,661	$56,889	$43,771	$39,889	$30,635,068	768
False Positives-Low Cost Users in 2011 to 2012	$11,444	$8,511	$0	$25,511	($17,000)	($8,823,000)	519
False Positives-High Cost Users in 2011 to 2012	$13,029	$46,551	$31,655	$31,896	$14,655	$5,085,204	347
AVERAGE PER PERSON WITH PROBABILITY OF 0.37 OR HIGHER	$55,711	$46,379	$28,944	$34,435	$11,944		
TOTAL FOR PERSONS WITH PROBABILITY OF 0.37 OR HIGHER	$105,237,222	$87,609,714	$54,675,037	$65,046,909		$22,562,037	1,889
PERSONS IN 10TH COST DECILE IN 2010 WITH TRIAGE TOOL PROBABILITY BELOW 0.37							
False Negatives-Low Cost Users in 2011 to 2012	$50,723	$8,067	$0	$25,067	($17,000)	$105,237,222	1,428
False Negatives-High Cost Users in 2011-2012	$72,621	$60,115	$40,878	$36,237	$23,878	$21,561,878	903
AVERAGE PER 10TH DECILE PERSON WITH PROBABILITY BELOW 0.37	$59,206	$28,230	$15,836	$29,394	($1,164)		
TOTAL FOR 10TH DECILE PERSONS WITH PROBABILITY BELOW 0.37						($2,714,122)	2,331

The bottom part of *Table 6* presents the results for *false negatives,* who were scored low for 2010 because of low service utilization levels recorded in 2008 and 2009. They were *false negatives* because service levels increased significantly in 2010, pushing them into the top decile of high cost users. However, 61 percent of *false negatives* had a cost spike in 2010 and very low cost levels in 2011. Based on actual cost outcomes in 2011 they belong in the *true negative* group.

The remaining 39 percent of *false negatives* showed higher costs in 2011, and if they could be identified and housed they would also contribute to significant cost savings – over $21 million. However, taken altogether, *false negatives* yield negative cost outcomes because of the higher share of individuals with low future costs. While the model fails to score this group high enough to be included in the high-cost group because of their low pre-2010 service levels, we see that if they had been housed, costs per individual would have exceeded savings by over $1,000.

On the other hand, the 903 individuals who were high–cost *false negatives* in 2010 but had higher costs in 2011, would receive a higher probability score when they were screened again in 2011 because of their rising service levels and would be included in the high–cost group as a result of this subsequent rescreening. Hence, 42 percent of *false negatives* are projected to become *true positives* the next year, contributing to significant cost savings when re-screened and placed in permanent supportive housing.

The 2011 cost analysis for 1,000 persons in our test population with the highest probability scores, scores at or above 0.53, is shown in *Table 7*. Almost two–thirds (653 individuals) were *true positives*. Evaluating actual 2011 costs we observe that 122 of them became low-cost users in 2011, while more than half, 531, remained high-cost users. On the other hand, 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. Applying the 68 percent anticipated cost savings rate to this group, we show annual cost savings per individual in the fourth column. As described above for *Table 6*, we present adjusted costs, the net savings per individual and total savings for the group in columns five through seven.

Based on our assumptions, we estimate saving of almost $19 million in 2011 if the top 1,000 high-probability service users were housed permanently with supportive services. 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 7: Cost Savings for 2011 at the Cut–off Level of 0.53

Probability-Cost Group	YEAR PREDICTED 2010 Average Cost per Person if Unhoused	1 YEAR AFTER PREDICTION 2011 Average Cost per Person if Unhoused	POST-HOUSING COSTS AND SAVINGS 2011 Cost Savings per Person Housed (68% for 10th decile persons)	2011 Adjusted Cost per Person Housed ($17,000 added for housing cost)	2011 Net Savings per Person Housed	2011 Total Savings for Group	Number of Cases
PERSONS WITH TRIAGE TOOL PROBABILTY OF 0.53 OR HIGHER							
True Positives-Low Cost Users in 2011 to 2012	$111,580	$11,496	$0	$28,496	($17,000)	($2,074,000)	122
True Positives-High Cost Users in 2011 to 2012	$96,892	$86,947	$59,124	$44,823	$42,124	$22,367,823	531
False Positives-Low Cost Users in 2011 to 2012	$12,427	$8,829	$0	$25,829	($17,000)	($3,094,000)	182
False Positives-High Cost Users in 2011 to 2012	$13,579	$43,560	$29,621	$30,939	$12,621	$2,082,432	165
AVERAGE COST PER PERSON WITH PROBABILITY OF 0.53 OR HIGHER	**$69,565**	**$56,366**	**$36,282**	**$37,083**	**$19,282**		
TOTAL FOR PERSONS WITH PROBABILITY OF 0.53 OR HIGHER	**$69,564,661**	**$56,365,647**	**$36,282,255**	**$37,083,392**		**$19,282,255**	**1,000**
PERSONS IN THE 10TH COST DECILE IN 2010 WITH TRIAGE TOOL PROBABILITY BELOW 0.53							
False Negatives-Low Cost Users in 2011 to 2012	$52,544	$8,262	$0	$25,262	($17,000)	($26,537,000)	1,561
False Negatives-High Cost Users in 2011-2012	$75,177	$63,479	$43,166	$37,313	$26,166	$29,828,921	1,140
AVERAGE COST PER 10TH DECILE PERSON WITH PROBABILITY BELOW 0.53	**$62,097**	**$31,567**	**$18,219**	**$30,348**	**$1,219**		
TOTAL FOR 10TH DECILE PERSONS WITH PROBABILITIES BELOW 0.53						**$3,291,921**	**2,701**

Similar to *Table 6,* the bottom part of *Table 7* presents the results for *false negatives,* who were scored low for 2010 because of low service utilization levels recorded in 2008 and 2009, but then moved to the top decile of high cost users in 2010 because of a significant increase in their service levels. Fifty-eight percent of *false negatives* had a cost spike in 2010 reaching $52,000. However, since their cost levels dropped to $8,000 in 2011, they belong in the *true negative* group.

The remaining 42 percent of *false negatives* showed stable high costs both in 2010 and in 2011, and if they could be identified and housed they would also contribute to significant cost savings – almost $30 million. However, as in *Table 6*, because of the large share of individuals with lower future costs, if all *false negatives* had been housed, savings per individual would be merely $1,200.

Cost Savings in 2012

The cost results in 2012 for individuals with a probability score of 0.37 or higher are shown in *Table 8*. The results show that cost levels in 2012 are lower than in 2010 or 2011 due to *regression to the mean*. This led to lower cost savings in 2012, with reduced county spending as a result of the recession being an additional downward factor. However, we still estimate almost $16 million of savings in the third year, which corresponds to over $8,000 saved per housed individual. The cumulative savings for 2011 and 2012 exceeds $38 million.

Table 8: Cost Savings for 2012 at the Cut-off Level of 0.37

Probability-Cost Group	YEAR PREDICTED 2010 Average Cost per Person if Unhoused	2 YEARS AFTER PREDICTION 2012 Average Cost per Person if Unhoused	POST-HOUSING COSTS AND SAVINGS 2012 Cost Savings per Person Housed (68% for 10th decile persons)	2012 Adjusted Cost per Person Housed ($17,000 added for housing cost)	2012 Net Savings per Person Housed	2012 Total Savings for Group	Number of Cases
PERSONS WITH TRIAGE TOOL PROBABILTY OF 0.37 OR HIGHER							
True Positives-Low Cost Users in 2011 to 2012	$90,989	$8,449	$0	$25,449	($17,000)	($4,335,000)	255
True Positives-High Cost Users in 2011 to 2012	$93,196	$67,826	$46,122	$38,704	$29,122	$22,365,450	768
False Positives-Low Cost Users in 2011 to 2012	$11,444	$7,373	$0	$24,373	($17,000)	($8,823,000)	519
False Positives-High Cost Users in 2011 to 2012	$13,029	$52,900	$35,972	$33,928	$18,972	$6,583,284	347
AVERAGE PER PERSON WITH PROBABILITY OF 0.37 OR HIGHER	$55,711	$40,459	$25,359	$32,100	$8,359		
TOTAL FOR PERSONS WITH PROBABILITY OF 0.37 OR HIGHER	$105,237,222	$76,427,750	$47,903,734	$60,637,016		$15,790,734	1,889
PERSONS IN 10TH COST DECILE IN 2010 WITH TRIAGE TOOL PROBABILITY BELOW 0.37							
False Negatives-Low Cost Users in 2011 to 2012	$50,723	$6,115	$0	$23,115	($17,000)	($24,276,000)	1,428
False Negatives-High Cost Users in 2011-2012	$72,621	$45,250	$30,770	$31,480	$13,770	$12,434,310	903
AVERAGE PER 10TH DECILE PERSON WITH PROBABILITY BELOW 0.37	$59,206	$21,275	$11,920	$26,355	($5,080)		
TOTAL FOR 10TH DECILE PERSONS WITH PROBABILITY BELOW 0.37						($11,841,690)	2,331

The projected savings for the 39 percent of *false negatives* who have become *true positives* because of their high costs would add almost $12 million in net savings in 2012 if they were re-screened. However, the other 61 percent of *false negatives* have low costs and if they had been housed there would have been an estimated net loss of over $24 million in public funds. Hence, the probabilities produced by the triage tool avoid a significant loss by correctly excluding the *false negative* cases, because most have low long-term costs.

The cost results in 2012 for the 1,000 individuals with a probability score of 0.53 or higher are shown in *Table 9*. The higher cut-off level results in higher savings per person. We estimate over $16 million net savings for this group in the third year if they had been housed, which corresponds to over $16,000 saved per housed individual. The cumulative savings for 2011 and 2012 exceeds $35 million.

The projected savings for the 42 percent *false negatives* in this group who became *true positives* because of their long-term high costs would have yielded almost $17 million in net savings in 2012 if they had been re-screened and housed. However the other 58 percent of *false negatives* had low-long-term costs and if they had been housed there would have been an estimated net loss of over $26 million in public funds. As with the population shown in *Table 8*, the probabilities produced by the triage tool in *Table 9* avoid a significant loss by correctly excluding the *false negative* cases, unless their cost trajectories are disaggregated through re-screening, because most have low long-term costs.

Table 9: Cost Savings for 2012 at the Cut-off Level of 0.53

	YEAR PREDICTED	2 YEARS AFTER PREDICTION	POST-HOUSING COSTS AND SAVINGS				
Probability-Cost Group	**2010 Average Cost per Person if Unhoused**	**2012 Average Cost per Person if Unhoused**	**2012 Cost Savings per Person Housed (68% for 10th decile persons)**	**2012 Adjusted Cost per Person Housed ($17,000 added for housing cost)**	**2012 Net Savings per Person Housed**	**2012 Total Savings for Group**	**Number of Cases**
PERSONS WITH TRIAGE TOOL PROBABILTY OF 0.53 OR HIGHER							
True Positives-Low Cost Users in 2011 to 2012	$111,580	$9,419	$0	$26,419	($17,000)	($2,074,000)	122
True Positives-High Cost Users in 2011 to 2012	$96,892	$74,546	$50,691	$40,855	$33,691	$17,890,070	531
False Positives-Low Cost Users in 2011 to 2012	$12,427	$7,648	$0	$24,648	($17,000)	($3,094,000)	182
False Positives-High Cost Users in 2011 to 2012	$13,579	$58,101	$39,509	$35,592	$22,509	$3,713,932	165
AVERAGE COST PER PERSON WITH PROBABILITY OF 0.53 OR HIGHER	**$69,565**	**$51,712**	**$33,436**	**$35,276**	**$16,436**		
TOTAL FOR PERSONS WITH PROBABILITY OF 0.53 OR HIGHER	**$69,564,661**	**$51,711,645**	**$33,436,002**	**$35,275,643**		**$16,436,002**	**1,000**
PERSONS IN THE 10TH COST DECILE IN 2010 WITH TRIAGE TOOL PROBABILITY BELOW 0.53							
False Negatives-Low Cost Users in 2011 to 2012	$52,544	$6,238	$0	$23,238	($17,000)	($26,537,000)	1,561
False Negatives-High Cost Users in 2011-2012	$75,177	$46,813	$31,833	$31,980	$14,833	$16,909,438	1,140
AVERAGE COST PER 10TH DECILE PERSON WITH PROBABILITY BELOW 0.53	**$62,097**	**$23,363**	**$13,436**	**$26,928**	**($3,564)**		
TOTAL FOR 10TH DECILE PERSONS WITH PROBABILITIES BELOW 0.53						**($9,627,562)**	**2,701**

Conclusion

We selected 0.37 and 0.53 as the cut-off levels for this cost analysis, but 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. We estimated that if a cut-off level of 0.20 is selected, over 4,000 homeless individuals would be identified and the results would be roughly break-even, with expenditures matching savings.

Over the three 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).

Using the Silicon Valley Triage Tool

Introduction

This chapter provides operational information for using the Silicon Valley Triage Tool. The tool builds on experience from Los Angeles in developing predictive analytic screening tools as well as operational experience using the tools to identify and house high-cost residents experiencing homelessness. It benefits from that experience as well as the much larger and better quality body of data that was available from Santa Clara County.

Comparative Performance of Los Angeles and Silicon Valley Triage Tools

Two triage tools have been developed in Los Angeles. Tool #1 uses both health care and justice system data. Tool #2 was developed to use just data that is available in hospitals. It makes more extensive use of diagnostic data but does not use justice system data.

Both Los Angeles tools were limited by the two-year time window of data for homeless residents, as well as a smaller data sample. Because of the two-year window, the Los Angeles tools predict who currently is in the highest cost group, whereas the Silicon Valley tool predicts who will be in the high cost group in the coming year.

The predictive function of the Silicon Valley tool is valuable because future costs are a key consideration in deciding who should have priority access to permanent supportive housing.

The predictive performance of the Silicon Valley Triage tool was compared to the performance of the two 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 two measures of performance.

The first measure, shown in *Figure 15*, is the proportion of high-cost homeless persons correctly identified by each model. All three tools identify the most high-cost individuals when the probability threshold is lowest, thereby encompassing the largest population. However, as shown in Figure 16, lower probability thresholds result in less accurate predictions. When applied to Los Angeles data, the Silicon Valley tool is slightly less accurate than Los Angeles tool #1 and slightly more accurate than Los Angeles Tool #2 in identifying everyone in the high-cost group. When applied to Santa Clara data it is much more accurate than either Los Angeles tool.

Figure 15: Percent of High-Cost Homeless Residents of Los Angeles and Santa Clara Counties Correctly Identified by Each Triage Tool
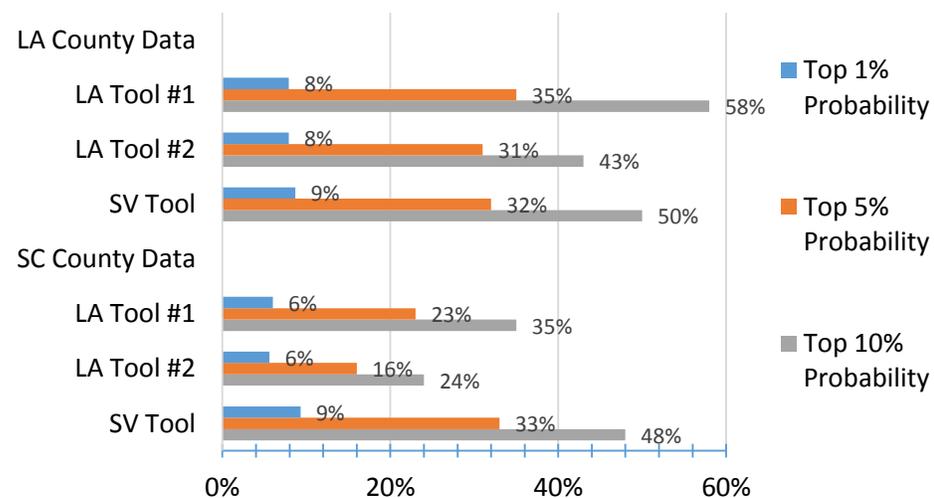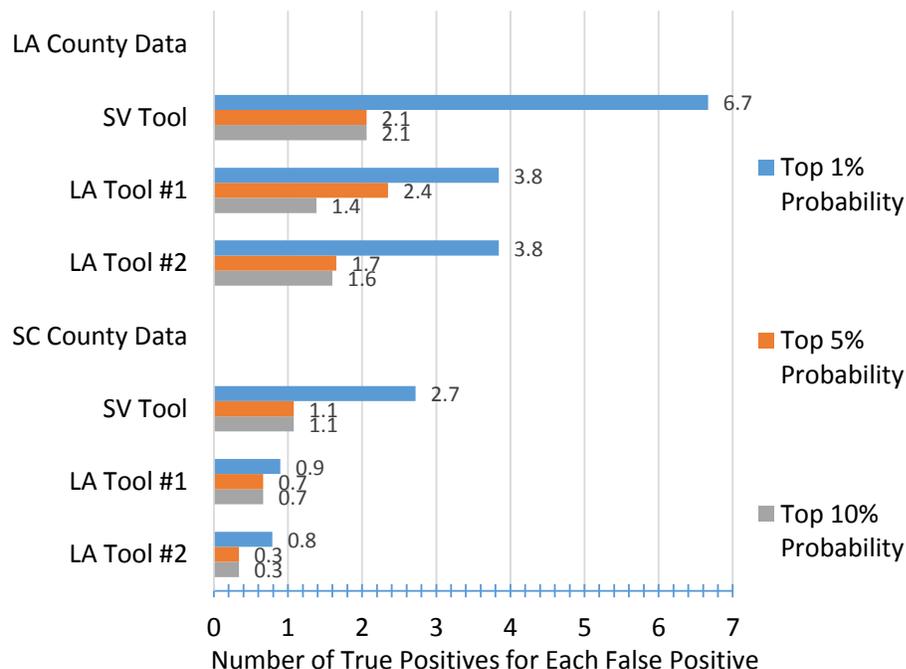
Figure 16: Ratio of True Positives to False Positives among Homeless Residents of Los Angeles and Santa Clara Counties Identified by Each Triage Tool as being in the High-Cost Group

The second measure, shown in *Figure 16*, is the proportion of persons predicted to be high-cost homeless who truly were high-cost persons. These predictions become more accurate as the probability threshold is raised. Measured against this benchmark when applied to Los Angeles data for individuals that each tool ranked as being among the top 5 percent and 10 percent with the greatest probability of being in the high-cost group, the Silicon Valley tool is roughly as accurate as Los Angeles tool #1 and more accurate than Los Angeles tool #2 in correctly differentiating high-cost individuals. It is much more accurate than either Los Angeles tool in correctly differentiating high-cost individuals in the top 1 percent probability group. When applied to Santa Clara County data, the Silicon Valley tool is much more accurate than either Los Angeles tool in correctly differentiating high-cost individuals in all three probability groups.

Based on both performance measures, the Silicon Valley tool demonstrates comparable or higher accuracy when run on Los Angeles data and much higher accuracy when applied to the Santa Clara data. This comparison is discussed in greater detail in *Section D* of the *Appendix*, and verifies that while the performance of the Silicon Valley Triage Tool on Santa Clara data is comparable to the performance of Los Angeles tools on Los Angeles data, its performance with out-of-county data is much better than the Los Angeles tools.

While the performance of the Silicon Valley Triage Tool on Santa Clara data is comparable to the performance of Los Angeles tools on Los Angeles data, its performance with out-of-county data is much better than the Los Angeles tools.

System-Based Tool

The Silicon Valley Tool, like the Los Angeles tools, can be used to screen cases individually or to screen entire data bases. And like the Los Angeles tools it is system-based, 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.

Santa Clara County will carrying out system-wide record linkages every year, making it feasible to screen large groups rather than just screening cases individually. This is the most efficient approach to screening and makes it possible to prioritize homeless residents for access to the available supply of permanent supportive housing based on their probability scores. Targeted individuals can be flagged in client databases so that housing can be offered to them the next time they are encountered.

Figure 17: User Interface with Triage Tool in Excel Format

Santa Clara County Triage Tool for Residents Experiencing Homelessness

GUIDE
37-100% Strong probability
35-36.9% Borderline
0-34.9% Below threshold

Follow the instructions in column "C" when the description in column "B" fits a client
Use one column per client to enter information for up to 25 clients in columns D to AB
Enter 1 or a number in rows 12 to 54 of each column as indicated by instructions in column "C"
The probability that a client is in the highest cost group is shown in row 8 of each column

PROBABILITY
2.2% 2.2%

Question Number	INFORMATION FOR TRIAGE TOOL	Instructions for Entering Values	Enter Values for Client 1	Enter Values for Client 2
	Constant value	Auto	Automatic	Automatic
	DEMOGRAPHICS			
1	Age 18-45	Enter 1 if true		
2	Age 46-65	Enter 1 if true		
3	Age 65+	Enter 1 if true		
4	Female	Enter 1 if true		
	CRIMINAL JUSTICE			
5	100+ days of probation in the last two years	Enter 1 if true		
6	Arrested in last two years	Enter 1 if true		
7	Jail booking in last two years	Enter 1 if true		
8	Jail security classification of 3 or 4 this year	Enter 1 if true		
9	Booking code 849A or 849B - arrested for inebriation and released within 48 hours - this year	Enter 1 if true		
10	*Number* of arrests this year	Enter *Number*		
11	*Number* of days in jail this year	Enter *Number*		
	MEDICAL DIAGNOSES			
12	Diagnosed with chronic medical condition *(see list of diagnoses on accompanying worksheet)*	Enter 1 if true		
13	Medical encounter with diagnosis of adjustment reaction ICD-9 309 this or last year	Enter 1 if true		
14	Medical encounter with diagnosis of heart disease ICD-9 401-429 this or last year	Enter 1 if true		
15	*Number* of medical encounters with ICD-9 (Organ Failures) 569-573, 576-578, 585-594, or 596 in last two years	Enter *Number*		
16	Medical encounter with diagnosis of schizophrenia ICD-9 295 in last two years	Enter 1 if true		
17	*Number* of medical encounters with diagnosis of neoplasm (ICD-9 140 to 239) in last two years	Enter *Number*		
18	Medical encounter with other ill-defined and unknown causes of morbidity and mortality (ICD-9 799) in last two years	Enter 1 if true		
19	Medical encounter with diagnosis of high-cost ICD-9 in last two years *(see list on accompanying worksheet)*	Enter 1 if true		
	HEALTH AND EMERGENCY SERVICES			
20	Emergency Medical Service (EMS) encounter this year	Enter 1 if true		
21	Emergency Medical Service (EMS) encounter last year	Enter 1 if true		
22	Two or more Emergency Medical Service (EMS) encounters in last two years	Enter 1 if true		
23	Admitted as hospital inpatient via emergency unit admission or transfer from psychiatric facility in last two years	Enter 1 if true		
24	Outpatient Psychiatric Emergency Services or ambulatory surgery this year	Enter 1 if true		
25	*Number* of hospital inpatient admissions this year	Enter *Number*		
26	*Number* of hospital inpatient days in last two years	Enter *Number*		
27	VMC non-inpatient (ER or clinic) health system encounter this year	Enter 1 if true		
28	*Number* of non-inpatient health system encounters (ER or clinic visits) this year	Enter *Number*		
29	Eleven or more non-inpatient (ER or clinic) health system encounters this year	Enter 1 if true		
	BEHAVIORAL HEALTH			
30	*Number* of Mental Health outpatient days in the last two years	Enter *Number*		
31	Two or more Mental Health outpatient visits in the last two years	Enter 1 if true		
32	*Number* of Mental Health inpatient admissions this year	Enter *Number*		
33	Two or more Mental Health inpatient admission in the last two years	Enter 1 if true		
34	Substance abuse indicated by any recorded medical diagnosis or justice system charge *(see lists on accompanying worksheets)*	Enter 1 if true		
35	*Number* of DADS encounters in the last two years	Enter *Number*		
	HOMELESS MANAGEMENT INFORMATION SYSTEM AND SOCIAL SERVICES			
36	Chronic homeless flag in HMIS record	Enter 1 if true		
37	Public assistance benefits received this year	Enter 1 if true		
38	Two or more months of food stamp payments in the past two years	Enter 1 if true		
	INTERACTIONS BETWEEN DATA ALREADY ENTERED			
	Age 18-45 & 10+ Mental Health outpatient visits 2+ last 2 years	Auto	Automatic	Automatic
	Age 46-65 & 10+ Mental Health outpatient visits 2+ last 2 years	Auto	Automatic	Automatic
	Age 65+ & 10+ Mental Health outpatient visits 2+ last 2 years	Auto	Automatic	Automatic
	Medical encounter with diagnosis of heart disease ICD-9 402-429 this or last year & 10+ Mental Health outpatient visits > 10 last 2 years	Auto	Automatic	Automatic
	Emergency unit admission or transfer from psychiatric facility & Mental Health inpatient admissions > 1 this or last year	Auto	Automatic	Automatic
	Days in jail this year IF arrested this year	Auto	Automatic	Automatic
	Chronically homeless in HMIS & diagnosed ICD-9 309 (adjustment reaction)	Auto	Automatic	Automatic
	Diagnosed with chronic medical condition & Current year emergency unit admission or transfer from psychiatric facility via EMS	Auto	Automatic	Automatic
	Diagnosis of chronic medical condition & Current year emergency unit admission or transfer from psychiatric facility via VMC outpatient unit	Auto	Automatic	Automatic
	Arrested and booked Cell blocks=849A , 849B in past year & Emergency unit admission or transfer from psychiatric facility	Auto	Automatic	Automatic
	100+ probation days in current year & Booking code 849A or 849B in current year - arrested for inebriation and released within 48 hours	Auto	Automatic	Automatic

The software code for the tool can be downloaded at no cost from the Economic Roundtable web site, www.economicrt.org, by agencies that want to screen entire client databases. The tool has also be exported into Excel and can be downloaded in that format from the Roundtable web site. The tool as it appears in Excel format is shown in *Figure 17*. The statistical formulas that produce triage tool probabilities are located below the user interface and should not be disturbed.
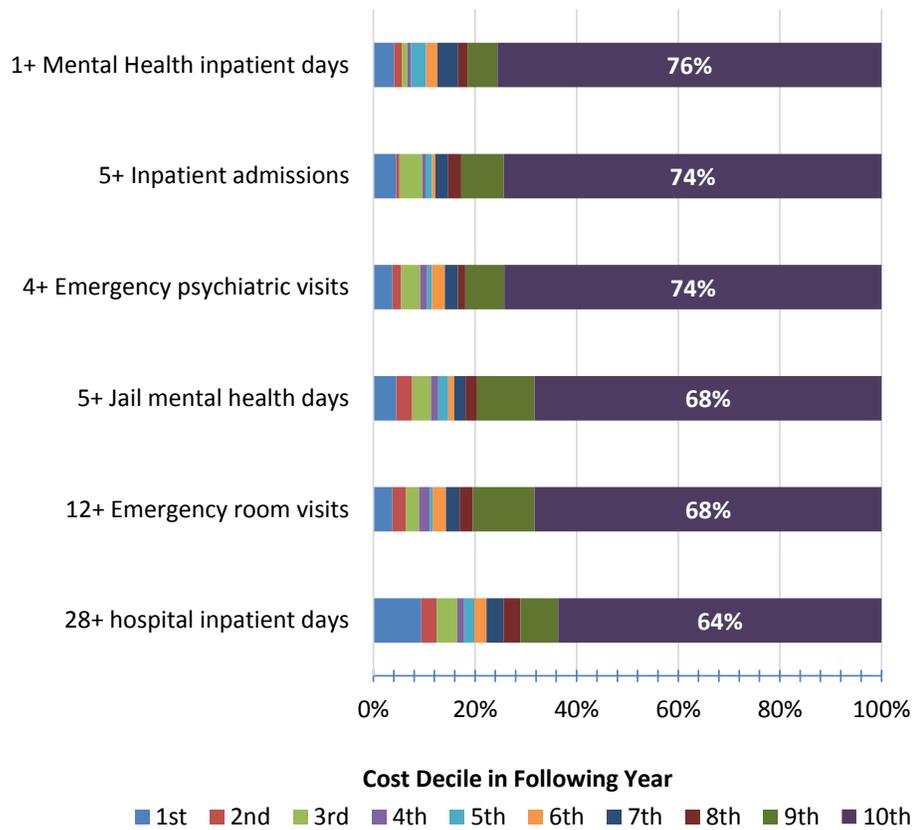
Screening Clients Individually

The tool can also be used to screen clients individually, using a version of the tool that has been exported into Excel.

A data collection form for assembling the 38 pieces of information needed to screen clients individually using the tool is provided in the *Appendix*. It begins by asking for the name, birthdate and place of birth of the client. This is needed for the triage tool question about age. Other information includes:

1. *Eligibility*: Client's homeless status and whether his or her background includes something that will prevent them from getting access to subsidized housing (these barriers vary by program and locality).

2. *General Information*: Demographics and information about whether the client will be able to live independently in permanent supportive housing or needs a higher level of care such as a skilled nursing facility.

3. *Justice System History*: Information about justice system involvement over the past two years that is needed for the triage tool.

4. *Diagnostic Information*: This is a check-off list for any diagnoses that are present in the client's medical record. The three columns on the right side of the table indicate how the information should be used. The third-from-the-right column indicates which of the diagnoses are associated with high costs. This information is needed to answer question 19. A complete list of high-cost diagnoses is provided on a worksheet in the triage tool Excel file. The second-from-the-right column indicates which of the diagnoses are associated with a chronic medical condition. This information is needed to answer question 12. A complete list of chronic medical conditions is provided on a worksheet in the Excel file. The last column on the right indicates which questions correspond with each diagnostic category.

5. *Health and Emergency Services:* Information about encounters with different types health care providers over the past two years.

6. *Behavioral Health:* Information about encounters with different types of behavioral health care providers over the past two years.

7. *HMIS and Social Services:* Information about public assistance benefits and documentation of chronic homeless status by HUD-funded homeless service provider.

8. *Automatic Inclusion in Highest Cost Group*: If a client meets any of six service use benchmarks in the past two years, they can automatically be included in the high cost group. This is an alternative to entering data into the triage tool. These benchmarks are <u>not</u> valid as an alternative way of re-assessing cases that have been scored with the triage tool and have probabilities below the 0.37 threshold. Use of the benchmarks is explained in the following section.

Figure 18: Cost Decile Based on Service Use in Preceding Two Years

1+ Mental Health inpatient days	76%
5+ Inpatient admissions	74%
4+ Emergency psychiatric visits	74%
5+ Jail mental health days	68%
12+ Emergency room visits	68%
28+ hospital inpatient days	64%

Cost Decile in Following Year

■ 1st ■ 2nd ■ 3rd ■ 4th ■ 5th ■ 6th ■ 7th ■ 8th ■ 9th ■ 10th

Service Use Benchmarks

A small number of clients have levels of service use in the past two years that make it likely they are in the high-cost group based on a single benchmark. The six benchmarks and the percent of clients meeting those benchmarks who are in the 10th and highest cost decile are shown in *Figure 18*.

It is important to note that the triage tool identifies 71 percent of cases that meet any of these benchmarks as being in the highest cost group. If a case meets one of these benchmarks but receives triage tool probability below 0.37, it is unlikely that the individual will be in the high-cost group. The percent of cases that meet one of these benchmarks but has a triage tool score below 0.37 and yet is in the high-cost group is as follows:

- 5+ Hospital inpatient admissions   10 percent
- 4+ Emergency psychiatric visits   16 percent
- 5+ Jail mental health days   20 percent
- 1+ Mental Health inpatient days   27 percent
- 12+ Emergency room visits   27 percent
- 28+ hospital inpatient days   29 percent

A small number of clients have levels of service use in the past two years that make it likely they are in the high-cost group based on a single benchmark.

The service use benchmarks can be used instead of the triage tool. But the benchmarks should <u>not</u> be used to reassess cases that been scored using the triage tool and received a probability below the 0.37 cut-off level.

Information in the Excel Triage Tool File

The Excel file that contains the triage tool has a number of additional worksheets that provide information related to using the triage tool. The file includes:

1. *Triage Tool:* Instructions are provided for entering data. The tool provides one column per client for up to 25 clients at a time.

2. *High-Cost ICD-9 Diagnostic Codes:* A list of all 59 high-cost IDC-9 codes is provided on the second worksheet.

3. *Crosswalk from ICD-9 to ICD-10*: A crosswalk from ICD-9 codes to the new ICD-10 codes that will soon be in use is provided on the third worksheet.

4. *Chronic ICD-9 Codes*: A list of 13,472 ICD-9 diagnostic codes that are associated with chronic medical conditions is provided on the fourth worksheet.

5. *Substance Abuse ICD-9 Codes*: A list of 13 ICD-9 codes associated with substance abuse is provided on the fifth worksheet.

6. *Substance Abuse Statutes*: A list of 254 statutes in effect in cities in Santa Clara County that prohibit different behaviors associated with substance abuse is provided on the sixth worksheet.

7. *Notes*: Additional information about the file is provided on the seventh Worksheet.

Best Practices when Screening and Housing Individuals Experiencing Homelessness

Two central objectives in the screening process are first, to use correct and complete information to produce accurate predictions from the triage tool. Second, to defer telling clients about the possibility of being housed until after their eligibility has been established and their background has been reviewed to see if there are circumstances that are likely to preclude them from receiving housing subsidies. Six potential barriers to receive housing subsidies are listed on the first page of the *Triage Tool Screening Form for Homeless Residents* in the *Appendix* and include, for example, being an undocumented immigrant or a registered sex offender. The reason for waiting until after screening has been completed to tell clients about the prospect of housing is to avoid the possibility of demoralizing them by raising their hopes about having a place of their own in which to live, but then disappointing them when they are screened out of the project because they are not in the high-cost group or barred from receiving housing subsidies.

The threshold probability for identifying individuals in the high-cost group is a score of 0.37 or higher from the Silicon Valley Triage Tool. Whether the individual is identified when an entire client database is screened or through individuals screening, the next steps involve one-on-one assistance to immediately house the individual, obtain benefits and needed services, and assist the individual in renting and retaining a permanent supportive housing unit.

The key to helping high-need homeless individuals successfully transition into permanent housing is a housing navigator/case manager with extensive experience in building trust with clients and helping them obtain the benefits and services needed for their continued

> The threshold probability for identifying individuals in the high-cost group is a score of 0.37 or higher from the Silicon Valley Triage Tool.
> --------------------------

well-being. The transition from the hospital or jail to the navigator's care takes place through a warm handoff in which discharge planner briefs the navigator on the client's social and medical background, providing information about the individual's personal characteristics, history of hospital and/or jail use, presenting issues, diagnoses, and underlying problems. This is followed by a personal introduction of the navigator to the client.

The navigator's meeting with the client is the final step in the screening process and the first step in building a long-term relationship with growing trust. In order to obtain a lease from a permanent supportive housing provider, the client typically needs to have both a Section 8 housing voucher to pay the bulk of the rent and an ongoing source of income, most often from Supplemental Security Income (SSI) to cover the tenant's portion of the rent and to pay living expenses. Among other things, the interview explores further whether there are any showstoppers for obtaining a Section 8 housing voucher. In some instances project-based permanent supportive housing that is subsidized through a funding source other than Section 8 provides additional options for housing individual who have barriers to obtaining a Section 8 voucher.

After the navigator engages the client, the next steps and the necessary resources for helping the client include:

- Fulfillment of immediate needs such as filling *prescriptions* or providing *hygiene items*.
- Immediate *temporary housing*.
- Rapid connection with *health services* at Federally Qualified Health Centers (FQHCs).
- Rapid connection with mental health and behavioral health services when needed.
- Assistance in qualifying for *benefits* including Supplemental Security Income (SSI), Medicaid, and Section 8 housing vouchers.
- *Permanent supportive housing* as quickly as possible.
- *Ongoing engagement and support* after the client has leased a permanent supportive housing unit to help the individual become a successful tenant and retain his or her housing.

> When warranted, negative results from the triage tool should be overridden based on clinical judgment.

Clinical Override

The screening process should include an option to over-ride the triage tool probability score. If the probability is less than 0.37, the cut-off point, the reasonableness of this outcome should be open for review with medical staff. As explained earlier, the tool produces a significant number of *false negative* predictions. When warranted, negative results from the triage tool should be overridden based on clinical judgment that the individuals is likely to have continuing high public costs. If a client has recently been diagnosed with a high-cost medical condition, for example, this would be an important factor to consider in deciding whether to override a negative result from the triage tool and include the patient in the high-cost group that receives access to permanent supportive housing.

Conclusion

Performance of Model

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 eleven agencies administrative records. All these datasets include information that has an effect on our outcome of interest – becoming a high cost user next year. These factors include demographics, clinical variables and service utilization variables for the current and previous years as well as the cost of service data.

The algorithm we developed based on modern statistical and data-mining methods has very strong performance. And the model was validated using different cases than those used to develop the model. The model uses the strongest 38 predictors of becoming high-cost service users and is effective for targeting homeless individuals who should be given highest priority for access to permanent supportive housing.

As in all predictive modeling algorithms, this model is subject to the trade-offs between maximizing the number of high-cost homeless persons identified (*true positives*) and minimizing the number of low-cost homeless persons identified (*false positives*) at a given cut-off level. 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 tool is that it is based on service records over a five-year period. We have 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. The proportion of *false positives* is critical to the accuracy of the model as well as the cost efficiency of the intervention. A detailed analysis of this group over three years verified that many *false positives* become high-cost or close to high-cost users in the second year after the prediction. The remaining individuals who were incorrectly scored high represent the group with a one-time cost spikes during the prior two years, contributing to a false alarm in the year following the prediction. However, the proportion of this group is very low – only12.5 percent of the highest 1,000 probability scores.

We also analyzed *false negatives* over three years following the model prediction – homeless individuals who were not scored high by the model but had high cost levels in the prediction year. The analyses showed that a majority of this group are actually *true negatives* over the next two years because their high cost level in the scoring year represented a one-time cost spike.

Our analysis over three years not only verifies the Silicon Valley Triage Tool's strong performance in capturing high-cost service users, but linking and screening records annually also enables us to identify a portion of former *false negatives* who become high cost service users in the current year. The model identifies a significant number of current-year *false positives* and *false negatives* as *true positives* the next year while successfully excluding *true negatives*—those with one year cost spikes or stable low service levels.

Cost Results

We completed our validation by developing a business analysis to assess the cost effectiveness of the model. We identified 0.37 probability of being in the high-cost group as the optimal cut-off level. At this probability level we identify 5 percent of the population as the target group. We assessed potential 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.

Our cost analysis demonstrates that the model performs very efficiently by excluding *false negatives*, otherwise the net savings for this group would be negative over two years, costing more than the expected savings. Since many *false negatives* represent one-time cost spikes, their long-term service utilization is low.

We extended our cost analysis to different probability cut-off levels because the threshold can be raised or lowered depending on policy objectives and availability of housing. We showed that at higher probability thresholds, per capita savings increase because a higher proportion of high-cost users are targeted. 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.

Recommended Use

The Silicon Valley Tool is twice as accurate as earlier tools developed in Los Angeles and is the best tool to use for identifying individuals who are likely to have high future costs. It 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.

It is possible to develop a simpler iteration of this data-intensive tool that is paper- or desktop computer-based and can be used by housing providers themselves in their offices to score clients and identify those at high risk of becoming high cost service users next year. However, this is not a near-term option because efforts are not yet underway to develop a simplified version of this tool.

Because of the level of effort required to obtain and integrate the necessary data, the most efficient use of this 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 are encountered.

The screening process 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.

References

Allison, P. (2012). *Logistic regression Using SAS: Theory and Application, Second Edition*. Cary, NC: SAS Institute.

Ash, A.S.; Zhao,Y; Ellis, R.P and Kramer, M.S. (2001). "Finding Future High-Cost Cases: Comparing Prior Cost versus Diagnosis-Based Methods." *Health Services Research* **36**(6): pp. 194–206.

Billings, J.; Dixon, J.; Mijanovich, T. and Wennberg, D. (2006). "Case Finding for Patients at Risk of Readmission to Hospital: Development of Algorithm to Identify High-Risk Patients." *British Medical Journal* 333(7563): 327. doi:10.1136/bmj.38870.657917.AE.

Billings, J.; Georghiou,T.; Blunt,I. and Bardsley,M. (2013). "Choosing a Model to Predict Hospital Admission: An Observational Study of New Variants of Predictive Models for Case Finding." *BMJ* Open 3: e003352. doi:10.1136/bmjopen-2013-003352.

Burt, M.; Paerson, C.L and Pearson, C.L (2005). *Strategies for Preventing Homelessness*, U.S> department of Housing.

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. (2015). "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 Program" *Housing Policy Debate*, DOI:10.1080/10511482.2015.1060249.

Center for Healthcare Research and Transformation. (2010). "Health Care Cost Drivers: Chronic Disease, Comorbidity, and Health Risk Factors in the U.S. and Michigan." Retrieved from <http://www.chrt.org/publication/health-care-cost-drivers-chronic-disease-comorbidity-health-risk-factors-u-s-michigan.

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** (3): pp. 68-79.

Culhane, D. P.; Metreux, 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**: pp. 107–163.

De Ville, B and Neville, P. (2013). *Decision Trees for Analytics Using SAS Enterprise Miner*. Cary, NC: SAS Institute.

Economic Roundtable (2009). *Where We Sleep: The Costs of Housing and Homelessness in Los Angeles*.

Economic Roundtable (2011). *Crisis Indicator: Triage Tool for Identifying Homeless Adults in Crisis*.

Economic Roundtable (2012). *Hospital to Home: Triage Tool II for Identifying Homeless Hospital Patient in Crisis*.

Economic Roundtable (2013). *Getting Home: Outcomes from Housing High Cost Homeless Hospital Patients*.

Economic Roundtable (2015). *Home Not Found: The Cost of Homelessness in Silicon Valley*.

Efron, B.; Hastie, T.; Johnstone, I. and Tibshirani, R. (2004). "Least Angle Regression" *Annals of Statistics* **32** *(2): pp. 407–499.*

Fleishman, H.A. and Cohen, J.W. (2010). "Using Information on Clinical Conditions to Predict High-Cost Patients." *Health Services Research* **45** (2): pp. 532–52.

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** pp: 445–50.

Gonen, M. (2007). *Analyzing Receiver Operating Characteristics with SAS*. Cary, NC: SAS Press Series.

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

Moturu, A. T.; Johnson, W. G. and Liu, H. (2010). "Predicting Future High-Cost Patients: A Real-World Risk Modeling Application" *International Journal of Biomedical Engineering and Technology*. **3**(1/2): pp. 114–32.

Rais, S.; Nazerian, A.; Ardal, S.; Chechulin, Y.; Bains,N. and Malikov, K. (2013). "High Cost Users of Ontario's Healthcare Services." *Healthcare Policy* **9** (1): pp. 44–51.

Rog, D. J.; Marshall, T.; Dougherty, R. H.; Preethy, G.; Daniles. A.S.; Ghose, S. S. and Delphin-Rittmon, M. E (2004). "Permanent Supportive Housing: Assessing the Evidence" *Psychiatric Services* **65** (3): pp. 287-294.

SadowskiL. 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** pp: 1771–78.

Sarma, K. S. (2013). Predictive Modeling with SAS Enterprise Miner: Practical solutions for Business Applications, Second Edition

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

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** (S2): pp. S324-S330.

Thomas, M. L.; Shears, J.K.;Pate, C. M. and Priester, M. A. (2014). *Moore Place Permanent Supportive Housing Evaluation Study*. University of North Carolina at Charlotte, Department of Social Work.

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*. County of Los Angeles, CEO, SIB.

Appendix

Appendix Contents

A. List of Potential Factors to Predict High Cost Homeless Persons

Table A-1: List of Potential Factors to Predict High Cost Homeless Persons

Demographic Factors
Age
Gender
Ethnicity
Language
Homelessness Factors
Chronic Homelessness
Duration of Homelessness
Number of CTA services
HMIS chronic homelessness indicator
Health Diagnoses
Total number of diagnostics
Chronic medical condition flag
001-139 Infections and Parasitic
042 HIV Disease
140-239 Neoplasms
240-279 Endocrine and Metabolic and Immune
250 Diabetes
280-289 Blood and Blood Organs
290-319 Mental Health Disorders
291 Alcohol-induced Mental Illness
292 Drug-induced Mental Illness
295 Schizophrenic Disorders
296 Episodic Mood Disorders
298 Other Nonorganic Psychoses
311 Depressive Disorders
309 Adjustment Reaction
303-5 Alcohol and drug dependence
320-389 Nervous System
390-459 Circulatory System
402-429 Heart Disease
451-459 Vein and lymphatic Disease
460-519 Respiratory System
470-478 Other Disease-Upper Respiratory Tract
480-488 Pneumonia and Influenza
490-496 Chronic Pulmonary Disease
520-579 Digestive System
569-73 and 76-78 and 85-94 and 96 liver, pancreas, intestines and kidneys
580-629 Genitourinary System
590-599, 614-616 Urinary Disease
680-709 Skin and Subcutaneous

710-739 Musculoskeletal System
780-799 Ill-defined Conditions
799 Other ill-defined and unknown causes of morbidity and mortality
800-999 Injury and Poisoning
V01-V89 Factors Influencing Health
Health and Emergency Services
If any Emergency Medical Service (EMS) encounters
Number of EMS encounters
If any hospital inpatient admissions
Admitted as hospital inpatient via emergency unit admission or transfer from psychiatric facility
Number of hospital inpatient admissions
Number of hospital inpatient days
If any outpatient visit with homeless targeted services
If any outpatient visit with mental health targeted services
VMC non-inpatient (ER or clinic) health system encounters
Number of non-inpatient health system encounters (ER or clinic visits)
Outpatient Psychiatric Emergency Services or ambulatory surgery
Outpatient services in VHC facility
Outpatient services in VSC facility
Behavioral Health Services
If any Mental Health inpatient admission
Number of Mental Health inpatient admissions this year
If any acute care Mental Health inpatient admission
Number of acute care Mental Health inpatient admissions this year
If any Mental Health outpatient admission
Number of Mental Health MH outpatient days
Substance abuse indicated by medical diagnosis or justice system charge
If any DADs encounter
Number of DADS encounters
Number of DA dosing (met and heroine detox) treatments
Criminal Justice including Custodial Health
If in probation
Days of probation
If arrested
If any jail booking
Jail security classification of 3 or 4
Booking code 849A or 849B - arrested for inebriation and released within 48 hours
Number of arrests
Number of days in jail
If housed in medical facilities
If housed in mental health facility
If any custodial health visit

If any custodial mental health visit
Social services
If any public assistance benefits received
If any General assistance benefits received
If any food stamp payments received
Number of months in public assistance
Number of months in General assistance
Number of months in Food Stamps

B. Averages of Model Variables for High Cost and Other Homeless Persons of the Validation Sample

Table A-2: Averages of Model Variables for High Cost and Other Homeless Persons of the Validation Sample

Variable	High Cost	Other
Demographics		
Age Less than 18 (Percent)	5%	10%
Age 18-45 (Percent)	56%	55%
Age 46-65 (Percent)	36%	31%
Age 65+ (Percent)	3%	4%
Female (Percent)	42%	54%
Criminal Justice		
100+ days of probation in the last two years (Percent)	18%	5%
Arrested in last two years (Percent)	46%	16%
Jail booking in last two years (Percent)	23%	9%
Jail security classification of 3 or 4 this year (Percent)	10%	1%
Booking code 849A or 849B - arrested for inebriation and released within 48 hours - this year (Percent)	8%	1%
Number of arrests this year (Mean)	.78	.16
Number of days in jail this year (Mean)	32.9	5.2
Health Diagnoses (all values are percent)		
Diagnosed with chronic medical condition	68%	35%
Medical encounter with diagnosis of adjustment reaction ICD-9 309 last 2 years	11%	3%
Medical encounter with diagnosis of heart disease ICD-9 401-429 this or last year	6%	2%
Number of medical encounters with ICD-9 (Organ Failures) 569-573, 576-578, 585-594, or 596 in last two years	.6	.1
Medical encounter with diagnosis of schizophrenia ICD-9 295 in last two years	14%	2%
Number of medical encounters with diagnosis of neoplasm (ICD-9 140 to 239) in last two years	.4	.1
Medical encounter with other ill-defined and unknown causes of morbidity and mortality (ICD-9 799) in last two years	17%	4%
Medical encounter with diagnosis of high-cost ICD-9 in last two years	52%	20%
Health and Emergency Services		
Emergency Medical Service (EMS) encounter this year (Percent)	30%	7%
Emergency Medical Service (EMS) encounter last year (Percent)	29%	7%

Variable	High Cost	Other
Two or more Emergency Medical Service (EMS) encounters in last two years	12%	1%
Admitted as hospital inpatient via emergency unit admission or transfer from psychiatric facility in last two years (Percent)	20%	4%
Outpatient Psychiatric Emergency Services or ambulatory surgery this year (%)	41%	15%
Number of hospital inpatient admissions this year (Mean)	.3	.06
Number of hospital inpatient days in last two years (Mean)	3.7	.6
VMC non-inpatient (ER or clinic) health system encounter this year (Percent)	68%	43%
Number of non-inpatient (ER or clinic visits) encounter this year (Mean)	6.2	2.3
Eleven + non-inpatient (ER or clinic) health system encounters this year (%)	20%	6%
Behavioral Health		
Number of Mental Health MH outpatient days in the last two years (Mean)	11.1	2.1
Two or more Mental Health outpatient visits in the last two years (Percent)	27%	9%
Number of Mental Health inpatient admissions this year (Percent)	17.6	1.2
Two or more Mental Health inpatient admission in the last two years (Percent)	20%	6%
Substance abuse indicated by any recorded medical diagnosis or justice system charge (Percent)	61%	31%
Number of DADS encounters in the last two years (Mean)	14.3	3.9
HMIS and Social Services		
Chronic homeless flag in HMIS record (Percent)	27%	11%
Public assistance benefits received in the current year (Percent)	46%	40%
Two or more months of food stamp payments in the past two years (Percent)	47%	44%

C. Logistic Regression Adjusted Odds Ratios and 95% Confidence Limits for Predictor Variables

Table A-3: Logistic Regression Adjusted Odds Ratios and 95% Confidence Limits for Predictor Variables

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 two years	1.15	1.03 – 1.28
Arrested in last two years	1.74	1.58 – 1.92
Jail booking in last two years	1.14	1.04 – 1.26
Jail security classification of 3 or 4 this year	1.63	1.41 – 1.89
Booking code 849A or 849B 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
Adjustment reaction this or last year	1.26	1.06 – 1.49
Heart disease this or last year	1.41	1.15 – 1.72
Number of Organ Failures in last two years	1.08	1.06 – 1.11
Schizophrenia in last two years	1.23	1.03 – 1.46
Number of neoplasm encounters in last two years	1.05	1.03 – 1.07
Other causes of morbidity and mortality in last two years	1.28	1.05 – 1.58
High-cost ICD-9 in last two years	1.12	1.009 – 1.24
Health and 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 two 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
VMC 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 MH outpatient days in the last two years	1.013	1.01 – 1.015
Two or more Mental Health outpatient visits in the last two years	1.4	1.23 – 1.59
Number of Mental Health inpatient admissions this year	1.002	1.002 – 1.003

Variable	Odds Ratio	95% Confidence Limits
Two or more Mental Health inpatient admission in the last two years	1.28	1.08 – 1.51
Substance abuse indicator	1.63	1.51 – 1.76
Number of DADS encounters in the last two years	1.002	1.002 – 1.002
HMIS and Social Services		
Chronic homeless flag in HMIS 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 in the past two years	.68	.59 - .79

D. Comparison of Silicon Valley Triage Tool to Earlier Tools Developed in Los Angeles

We presented the results of the Silicon Valley Triage Tool in Chapter III. While the results suggest that the performance of the tool is very strong, its performance relative to comparable models in identifying high-cost homeless persons is important to understand. In addition, it is important to assess the performance of the model using out-of-sample data, that is, data from another region, in comparison to other models because we often observe strong predictive power based on in-sample performance if the model over-fits a specific body of data, but poor out-of-sample performance when the model is applied to other data. Since a predictive model is intended to be applied to new data with unknown outcomes, cross-performance of the model is needed for validation.

However, we do not have any published results of comparable models developed for other jurisdictions with the exception of two triage tools built by Economic Roundtable in Los Angeles. These tools were developed by the Economic Round to identify the one-tenth of homeless individuals with the highest public costs (Economic Roundtable, 2011, 2012). Both tools were developed based on the Los Angeles County database created by linking service and cost records across county departments for a representative sample of almost 9,200 General Relief recipients with an episode of homelessness. Los Angeles Triage Tool 1 (LA 1) was developed using demographic, health care and criminal justice data fields. Los Angeles Triage Tool 2 (LA 2) was built based on demographic and health care data that is available to hospitals.

Here, we not only compare the performance of the Silicon Valley (SV) model to the Los Angeles models, we also validate both models by running them in cross data sets–SV on Los Angeles County data and LA1 and LA2 on Santa Clara County data. We show our results in *Table A-4* using the *sensitivity* and *positive predictive value* (*PPV*) measures. These values are presented for different scenarios–top 1 percent, 5 percent, 10 percent homeless persons with the highest predicted probability of becoming high-cost service users.

As described in Chapter III, 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 by assessing the proportions of *true positives* and *false negatives*(*true positives /(true positives + false negatives*)).

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 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 by assessing the proportions of *true positives* and *false positives* (*true positives / (true positives + false positives*)).

We show the two performance measures of 6 different model runs in *Table A-4*. These are the results of (1) SV model on Santa Clara data–as shown in Table 2 earlier; (2) SV model on Los Angeles data; (3) LA1 model on LA data; (4) LA2 model on LA data; (5) LA1 model on Santa Clara data and (6) LA2 model on Santa Clara data.

Table A-4: Model Comparison Results

Sample	SV-SC		SV-LA		LA1-LA		LA2-LA		LA1-SC		LA2-SC	
	Sensitivity	*PPV*	*Sensitivity*	*PPV*	*Sensitivity*	*PPV*	*Sensitivity*	*PPV*	*Sensitivity*	*PPV*	*Sensitivity*	*PPV*
Top 1%	9.3%	73%	8.7%	87%	7.9%	79%	7.9%	79%	6.0%	47%	5.6%	44%
Top 5%	33%	51%	32%	64%	35%	70%	31%	62%	23%	36%	16%	25%
Top 10%	48%	37%	50%	50%	58%	58%	43%	43%	35%	27%	24%	19%

The results show that, while the performance of the Silicon Valley Triage Tool on Santa Clara data is comparable to the performance of LA models on Los Angeles data, its out-of-sample performance is much better. The tool shows 9.3 percent sensitivity for the top 1 percent high-probability persons based on the Santa Clara data, which is higher than both LA models run on Los Angeles data, which report 7.9 percent. The *sensitivity* values of the Silicon Valley Triage Tool is comparable to the LA models for the top 5 percent and 10 percent samples. *PPV* values of LA models are higher, which is attributable to the much smaller population size of the Los Angeles data.

When we compare the out-of-sample performance of Silicon Valley and Los Angeles models, the results favor the Silicon Valley Triage Tool by a large margin. The tool achieved comparable or higher *sensitivity* and *PPV* values when we ran it on Los Angeles data. Its *PPV* values are also comparable or higher than the values reported by LA models on LA data. *PPV* values are particularly good, confirming that the proportions of true high-cost users identified by the tool are very high - 87 percent for the highest probability 1 percent and 64 percent for the highest 5 percent. These results verify that, when applied to the new data from another jurisdiction, the Silicon Valley Triage Tool achieves a similar or better performance.

In contrast, while LA models perform very well when applied to LA data, their out-of-performance are poor when applied to the Santa Clara data. In comparison, based on Santa Clara data, the Silicon Valley Triage Tool demonstrated 9.3 percent *sensitivity* and 73 percent *PPV* for the top 1 percent high-probability homeless persons, while LA models report 6 percent and 5.6 percent *sensitivity*, respectively, and 47 percent and 44 percent *PPV*, respectively. For the top 5 percent and 10 percent, results are also similar. The tool's *sensitivity* is 33 percent and *PPV* is 51 percent. In comparison, LA models show 23 percent and 16 percent *sensitivity*, respectively, and 36 percent and 25 percent *PPV* values, respectively.

This comparison provides a multi-regional benchmark for the Silicon Valley Triage Tool's performance and verifies that it demonstrates strong predictive performance when applied to data from another metropolitan region.

E. High Cost ICD-9 Medical Diagnostic Codes

Table A-5: High Cost ICD-9 Medical Diagnostic Codes

ICD-9 3-digit	ICD-9 Code and Description
012	Other respiratory tuberculosis
014	Tuberculosis of intestines, peritoneum, and mesenteric glands
015	Tuberculosis of bones and joints
018	Military tuberculosis
038	Septicemia
180	Malignant neoplasm of cervix uteri
193	Malignant neoplasm of thyroid gland
202	Other malignant neoplasms of lymphoid and histiocytic tissue
203	Multiple myeloma and immunoproliferative neoplasms
208	Leukemia of unspecified cell type
276	Disorders of fluid, electrolyte, and acid-base balance
285	Other and unspecified anemias
287	Purpura and other hemorrhagic conditions
295	Schizophrenic disorders
296	Episodic mood disorders
301	Personality disorders
305	Nondependent abuse of drugs
311	Depressive disorder, not elsewhere classified
312	Disturbance of conduct, not elsewhere classified
313	Disturbance of emotions specific to childhood and adolescence
314	Hyperkinetic syndrome of childhood
338	Pain, not elsewhere classified
356	Hereditary and idiopathic peripheral neuropathy
357	Inflammatory and toxic neuropathy
383	Mastoiditis and related conditions
403	Hypertensive chronic kidney disease
405	Secondary hypertension
414	Other forms of chronic ischemic heart disease
423	Other diseases of pericardium
424	Other diseases of endocardium
482	Other bacterial pneumonia
496	Chronic airway obstruction, not elsewhere classified
507	Pneumonitis due to solids and liquids
518	Other diseases of lung
532	Duodenal ulcer
560	Intestinal obstruction without mention of hernia
569	Other disorders of intestine
572	Liver abscess and sequelae of chronic liver disease
578	Gastrointestinal hemorrhage
585	Chronic kidney disease (CKD)
588	Disorders resulting from impaired renal function

ICD-9 3-digit	ICD-9 Code and Description
596	Other disorders of bladder
725	Polymyalgia rheumatica
751	Other congenital anomalies of digestive system
799	Other ill-defined and unknown causes of morbidity and mortality
863	Injury to gastrointestinal tract
867	Injury to pelvic organs
868	Injury to other intra-abdominal organs
885	Traumatic amputation of thumb (complete) (partial)
912	Superficial injury of shoulder and upper arm
916	Superficial injury of hip, thigh, leg, and ankle
924	Contusion of lower limb and of other and unspecified sites
935	Foreign body in mouth, esophagus, and stomach
959	Injury, other and unspecified
965	Poisoning by analgesics, antipyretics, and antirheumatics
970	Poisoning by central nervous system stimulants
971	Poisoning by drugs primarily affecting the autonomic nervous system
996	Complications peculiar to certain specified procedures
997	Complications affecting specified body systems, not elsewhere classified

F. Triage Tool Screening Information Form for Homeless Residents

Client Name: _____   Date of Birth:_____ Place of Birth:_____

Staff Name:_____   Today's Date:_____ Organization: _____

Staff Phone: _____   Client Location: _____

I. Eligibility

Is this client homeless?   ❑ **Yes**   ❑ **No**   ❑ **Don't Know**

A person who is homeless lacks a fixed, regular, and adequate nighttime residence; and has a primary nighttime residence that is:

- *A supervised publicly or privately operated shelter designed to provide temporary living accommodations (including welfare/voucher hotels or shelters, designed for homeless persons); or*
- *A public or private place not designed for, or ordinarily used as, a regular sleeping accommodation for human beings (street, park, hallway, freeway underpass, car)*

Individuals are <u>ineligible</u> for permanent supportive housing (and this pilot program) if any of the following attributes apply to them. Do any of the following describe this client?

1.	**Undocumented immigrant?**	❑ **yes**	❑ **no**
2.	**On parole for a violent crime?**	❑ **yes**	❑ **no**
3.	**Convicted of arson?**	❑ **yes**	❑ **no**
4.	**Convicted of operating a methamphetamine lab?**	❑ **yes**	❑ **no**
5.	**Committed an offense that requires registering as a sex offender?**	❑ **yes**	❑ **no**
6.	**Needs continuing nursing care?**	❑ **yes**	❑ **no**

If <u>none</u> of the above attributes apply to the client, please continue by providing the following information.

II. General Information

Gender?   ❑ **Male**
              ❑ **Female**

Ethnicity?   ❑ **African American**
                 ❑ **Asian American**
                 ❑ **European American**

Language?   ❑ **English**
                ❑ **Spanish**
                ❑ **Other** _____

                ❑ **Latino**
                ❑ **Pacific Islander**
                ❑ **Other** _____

Veteran?   ❑ **Yes**
               ❑ **No**
               ❑ **Don't Know**

Alcohol or drug dependency?   ❑ **Yes**
                                  ❑ **No**
                                  ❑ **Don't Know**

Does this client have a disability, that is, a physical or mental impairment that substantially limits one or more of the major life activities?   ❑ Yes   ❑ No   ❑ Don't Know

Is this client ambulatory?   ❑ Yes   ❑ No   ❑ Don't Know

Is this client expected to recover and live independently without continuing nursing care?

❑ Yes   ❑ No   ❑ Don't Know

Will this client need short-term respite care*?*   ❑ Yes   ❑ No   ❑ Don't Know

The Housing 1,000 Project places eligible high-need homeless clients temporarily in hotels, and then after the necessary documentation is assembled and approvals obtained, into permanent supportive housing. The project does not currently have access to respite care, recuperative care or skilled nursing beds. To be viable in the program, individuals must be able to live alone in a hotel and then in permanent supportive housing. Check (☑) if the client has any of the following conditions that are barriers to live in the housing that is available:

❑ *1. Wheel chair - assistance is not available to move clients into and out of wheel chairs. Clients in wheel chairs are viable for the program only if they are sufficiently ambulatory to be able to get out of the wheel chair, into a taxi, onto a toilet, and into a bed on their own.*

❑ *2. Colostomy bag*

❑ *3. Urinary catheter*

❑ *4. Tracheotomy*

❑ *5. Feeding tube*

❑ *6. Ongoing intravenous therapy*

❑ *7. Serious wounds that require ongoing wound care*

❑ *8. Incontinent of urine or feces*

❑ *9. Evidence of active tuberculosis*

❑ *10. At risk for alcohol withdrawal seizures/delirium tremors and needs medical detoxification*

❑ *11. Serious suicidal or assaultive risk*

Does this client have any of these barriers?   ❑ Yes   ❑ No   ❑ Don't Know

III. Justice System History

IN THE PAST 2 YEARS:

Provide the following information about justice system involvement in the past two years. None of this information excludes the client from possible housing referral.

On probation in past two years?   ❑ Yes   ❑ No   ❑ Don't Know

5.   If Yes, how many days?   _____ *days*

6. Arrested in past two years?   ❑ Yes   ❑ No   ❑ Don't Know

7. **Booked in jail in past two years?** ❑ Yes ❑ No ❑ Don't Know

8. **Jail security classification 3 or 4 this year?** ❑ Yes ❑ No ❑ Don't Know

9. **Booking code of 849A or B this year?** ❑ Yes ❑ No ❑ Don't Know

10. **Number of arrests this year?** _____ *arrests*

11. **Number of days in jail this year?** _____ *days in jail*

IV. Diagnostic Information

Instructions: Please review the client's medical records and check (☑) all diagnoses that apply. All high-cost diagnoses are flagged in the "High Cost" column. Many chronic diagnoses are flagged in the "Chronic" column, but there are additional medical conditions not shown on this list that are chronic. Consult the comprehensive list of chronic ICD-9 diagnoses that is included with the Excel version of the Silicon Valley Triage Tool if the client has medical conditions that are not listed but may be chronic. The triage tool question(s) that each diagnosis applies to are listed in the "Question" column.

12. **Diagnosed with chronic medical condition?** *(see "Chronic" column below and if the answer is still unclear, consult the list of chronic diagnoses in the worksheet with the Excel file of the triage tool)*

❑ Yes ❑ No ❑ Don't Know

☑	ICD-9	ICD-9 Description	High Cost	Chronic	Question
		INFECTIOUS AND PARASITIC DISEASES			
❑	012	Other respiratory tuberculosis	x		19
❑	014	Tuberculosis of intestines, peritoneum, and mesenteric glands			19
❑	015	Tuberculosis of bones and joints	x		19
❑	018	Miliary tuberculosis	x		19
❑	038	Septicemia	x		19
❑	140-239	NEOPLASM		x	17
❑	180	Malignant neoplasm of endocervix	x	X	19
❑	193	Malignant neoplasm of thyroid gland	x	x	19
❑	202	Other malignant neoplasms of lymphoid and histiocytic tissue	x	x	19
❑	203	Multiple myeloma and immunoproliferative neoplasms	x	x	19
❑	208	Leukemia of unspecified cell type	x	x	19
		ENDOCRINE, NUTRITIONAL, METABOLIC, IMMUNE DISORDERS			
❑	276	Disorders of fluid, electrolyte, and acid-base balance	x		19
❑	285	Other and unspecified anemias	x	x	19
❑	287	Purpura and other hemorrhagic conditions	x	x	19
	290-319	MENTAL DISORDERS			
❑	295	Schizophrenic disorders	x	x	16, 19
❑	296	Episodic mood disorders	x	x	19
❑	301	Personality disorders	x	x	19
❑	305	Nondependent abuse of drugs	x	x	19
❑	309	Adjustment reaction		x	13
❑	311	Depressive disorder, not elsewhere classified	x	x	19
❑	312	Disturbance of conduct, not elsewhere classified	x	x	19
❑	313	Disturbance of emotions specific to childhood and adolescence	x	x	19

☑	ICD-9	ICD-9 Description	High Cost	Chronic	Question
❑	314	Hyperkinetic syndrome of childhood	x	x	19
		DISEASES OF THE NERVOUS SYSTEM AND SENSE ORGANS			
❑	338	Pain, not elsewhere classified	x	x	19
❑	356	Hereditary and idiopathic peripheral neuropathy	x	x	19
❑	357	Inflammatory and toxic neuropathy	x	x	19
❑	383	Mastoiditis and related conditions	x		19
		DISEASES OF THE CIRCULATORY SYSTEM			
❑	401-429	Heart disease		x	14
❑	403	Hypertensive chronic kidney disease	x	x	19
❑	405	Secondary hypertension	x	x	19
❑	414	Other forms of chronic ischemic heart disease	x	x	19
❑	423	Other diseases of pericardium	x		19
❑	424	Other diseases of endocardium	x	x	19
		DISEASES OF THE RESPIRATORY SYSTEM			
❑	482	Other bacterial pneumonia	x		19
❑	496	Chronic airway obstruction, not elsewhere classified	x	x	19
❑	507	Pneumonitis due to solids and liquids	x		19
❑	518	Other diseases of lung	x		19
		DISEASES OF THE DIGESTIVE SYSTEM			
❑	532	Duodenal ulcer	x	x	19
❑	560	Intestinal obstruction without mention of hernia	x		19
❑	569	Other disorders of intestine	x		15, 19
❑	570	Acute and subacute necrosis of liver			15
❑	571	Chronic liver disease and cirrhosis		x	15
❑	572	Liver abscess and sequelae of chronic liver disease	x	x	15, 19
❑	573	Other disorders of liver		x	15
❑	576	Other disorders of biliary tract		x	15
❑	577	Diseases of pancreas			15
❑	578	Gastrointestinal hemorrhage	x		15, 19
		DISEASES OF THE GENITOURINARY SYSTEM			
❑	585	Chronic kidney disease (CKD)	x	X	15, 19
❑	586	Renal failure, unspecified		X	15
❑	587	Renal sclerosis, unspecified		X	15
❑	588	Disorders resulting from impaired renal function	x	X	15, 19
❑	589	Small kidney of unknown cause			15
❑	590	Infections of kidney			15
❑	591	Hydronephrosis			15
❑	592	Calculus of kidney and ureter			15
❑	593	Other disorders of kidney and ureter			15
❑	594	Calculus of lower urinary tract			15
❑	596	Other disorders of bladder	x	x	15, 19
		DISEASES OF THE MUSCULOSKELETAL SYSTEM			
❑	725	Polymyalgia rheumatica	x	x	19
		CONGENITAL ANOMALIES			
❑	751	Other congenital anomalies of digestive system	x	x	19
		SYMPTOMS, SIGNS, AND ILL-DEFINED CONDITIONS			
❑	799	Other ill-defined and unknown causes of morbidity and mortality	x		18, 19
		INJURY AND POISONING			

☑	ICD-9	ICD-9 Description	High Cost	Chronic	Question
❑	863	Injury to gastrointestinal tract	x		19
❑	867	Injury to pelvic organs	x		19
❑	868	Injury to other intra-abdominal organs	x		19
❑	885	Traumatic amputation of thumb (complete) (partial)	x		19
❑	912	Superficial injury of shoulder and upper arm	x		19
❑	916	Superficial injury of hip, thigh, leg, and ankle	x		19
❑	924	Contusion of lower limb and of other and unspecified sites	x		19
❑	935	Foreign body in mouth, esophagus, and stomach	x		19
❑	959	Injury, other and unspecified	x		19
❑	965	Poisoning by analgesics, antipyretics, and antirheumatics	x		19
❑	970	Poisoning by central nervous system stimulants	x		19
❑	971	Poisoning by drugs primarily affecting autonomic nervous sys.	x		19
❑	996	Complications peculiar to certain specified procedures	x		19
❑	997	Complications affecting specified body systems, NEC	x		19

13. to 19.   Answer triage tool questions 13-19 based on diagnostic information entered above.

V. Health and Emergency Services

Provide information about health and emergency services. None of this information excludes the client from possible housing referral.

20. **1+ Emergency Medical Service (EMS) encounters this year**   ❑ **Yes**   ❑ **No** ❑ **Don't Know**

21. **1+ Emergency Medical Service (EMS) encounters this year**   ❑ **Yes**   ❑ **No** ❑ **Don't Know**

22. **2+ Emergency Medical Service (EMS) encounters in last 2 years** ❑ **Yes**   ❑ **No** ❑ **Don't Know**

23. **Admitted as hospital inpatient via emergency unit admission or transfer from psychiatric facility in last two years**   ❑ **Yes**   ❑ **No** ❑ **Don't Know**

24 **Outpatient Psychiatric Emergency Services or ambulatory surgery this year**

                                                  ❑ **Yes**   ❑ **No** ❑ **Don't Know**

25 ***Number* of hospital inpatient admissions this year**   _____ *admissions*

26 ***Number* of hospital inpatient days in last two years**   _____ *inpatient days*

27 **VMC non-inpatient (ER or clinic) health system encounter this year**

                                                  ❑ **Yes**   ❑ **No** ❑ **Don't Know**

28 ***Number* of non-inpatient health system encounters (ER or clinic visits) this year**

                                                  _____ *ER or clinic visits*

29 **Eleven or more non-inpatient (ER or clinic) health system encounters this year**

                                                  ❑ **Yes**   ❑ **No** ❑ **Don't Know**

VI. Behavioral Health

Provide information about behavioral health services. None of this information excludes the client from possible housing referral.

30 ***Number* of Mental Health MH outpatient days in the last two years** _____
   outpatient days

31 **2+ Mental Health outpatient visits in the last two years** ☐ **Yes** ☐ **No** ☐ **Don't Know**

32 ***Number* of Mental Health inpatient admissions this year** _____ *admissions*

33 **2+ Mental Health inpatient admission in the last two years** ☐ **Yes** ☐ **No** ☐ **Don't Know**

34 **Substance abuse indicated by any recorded medical diagnosis or justice system charge** *(see list on triage tool worksheet)* ☐ **Yes** ☐ **No** ☐ **Don't Know**

35 ***Number* of DADS encounters in the last two years** _____ *encounters*

VII. HMIS and Social Services

Provide information about public assistance benefits and HUD-funded homeless services. None of this information excludes the client from possible housing referral.

36 **Chronic homeless flag in HMIS record** ☐ **Yes** ☐ **No** ☐ **Don't Know**

37 **Public assistance benefits received in the current year** ☐ **Yes** ☐ **No** ☐ **Don't Know**

38 **2+ months of food stamp payments in the past two years** ☐ **Yes** ☐ **No** ☐ **Don't Know**

VIII. Automatic Inclusion in Highest Cost Decile

If the client meets any one of the following six benchmarks for services used in the last two years, the individual can be automatically included in the highest cost decile without being scored using the triage tool.

Services Received in Last 2 Years:

1. 5 or more hospital inpatient admissions? ☐ **Yes** ☐ **No** ☐ **Don't Know**

2. 28 or more hospital inpatient days? ☐ **Yes** ☐ **No** ☐ **Don't Know**

3. 12 or more Emergency Room visits? ☐ **Yes** ☐ **No** ☐ **Don't Know**

4. 5 or more days jail mental health incarceration? ☐ **Yes** ☐ **No** ☐ **Don't Know**

5. 4 or more Psychiatric Emergency Service visits? ☐ **Yes** ☐ **No** ☐ **Don't Know**

6. 1 or more Mental Health inpatient days? ☐ **Yes** ☐ **No** ☐ **Don't Know**